# THE PROSODY OF BACKCHANNELS IN AMERICAN ENGLISH

*Stefan Benus* [1], *Agustín Gravano* [2], *Julia Hirschberg* [2]

[1] Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI, USA
[2] Department of Computer Science, Columbia University, New York, NY, USA
sb513@nyu.edu, agus@cs.columbia.edu, julia@cs.columbia.edu

## ABSTRACT

We examine prosodic and contextual factors characterizing the backchannel function of single affirmative words. Data is drawn from collaborative task-oriented dialogues between speakers of Standard American English. Despite high lexical variability, backchannels are prosodically well defined: they have higher pitch and intensity and greater pitch slope than affirmative words expressing other pragmatic functions. Additionally, we identify phrase-final rising pitch as a salient trigger for backchanneling.

**Keywords:** backchannels, prosody, English

## 1. INTRODUCTION

Backchannels such as *mmhm* and *okay,* which signal that the listener is attending to the speaker and does not wish to take the floor, are crucial for the synchronization of everyday communication and thus important to the development of spoken dialogue systems. For example, [5] found that backchannels (their 'continuers') comprise 19% of the dialogue acts in their corpus (a subset of the SWITCHBOARD corpus), second only to statements (36%). However, backchannels are also characterized by their ambiguity, since many lexical items can be employed as backchannels, and most of these items are themselves highly ambiguous. For example, *okay*, like most affirmative words, can mark a topic shift as well as conveying affirmation or a backchannel.

While the prosody of backchannels or its context might help to disambiguate backchannel uses from other discourse functions, relatively little is known about the phonetic characteristics of English backchannels or about the environment in which they occur. In a descriptive study of a small corpus, [4] identified a pitch contour that rises on the second syllable of *okay* and *uhhuh* as signaling a backchannel function. In another corpus study, [6] found that the most reliable prosodic cue preceding a backchannel was a region of low pitch. However, there has so far been little attempt to compare the discourse function, prosodic form, and context of affirmative words in general to see how the backchannel function in particular might be disambiguated.

In this study we investigate the prosodic and acoustic characteristics of backchannels in Standard American English. In Section 2 we describe the corpus of spoken dialogues, the discourse functions of affirmative words and our labeling scheme for them, and the prosodic and contextual features we examined. Section 3 describes our analyses and results. Section 4 summarizes the results and discusses their implications for Spoken Dialogue Systems.

## 2. THE CORPUS

The material for our study comes from the Columbia Games Corpus, a collection of 12 dyadic spontaneous task-oriented conversations elicited from speakers of St. American English. Subjects were paid to play two types of collaborative games (CARDS and OBJECTS) on laptops while seated in a soundproof booth divided by a curtain to ensure that all communication was verbal.

In the CARDS games, subjects received points for finding cards depicting the same objects on their different screens. One player described a card on her board, and the other searched for a full or partial match on his board. In the OBJECTS games, one player described the position of a target object with respect to other fixed objects on her screen, while the other tried to move his representation of the target object to the same position on his own screen. Points were given based on the proximity of the target object to its correct location. Both games were designed to encourage discussion, and the subjects switched roles repeatedly.

There were 13 subjects (7 males and 6 females); 11 played with two different partners in two different sessions and 2 played a single session. On average, each session took 45m 39s, totaling 9h 8m of dialogue for the whole corpus. All interactions were recorded, digitized, and downsampled to 16K. The recordings were orthographically

transcribed, and words were aligned to the source by hand. Nearly all of the OBJECTS part of the corpus has been intonationally transcribed, using the ToBI conventions ([1]).

### 2.1. Labeling discourse functions

We asked three labelers to independently classify all occurrences of single affirmative words {*alright, mmhm, okay, right, uhhuh, yeah, yep, yes, yup*} in the entire GAMES Corpus into one of 11 categories shown in Table 1.

**Table 1:** Labeled discourse functions

| | | | |
|---|---|---|---|
| **A1** | Acknowledgment / agreement | **K** | Question |
| **A2** | Backchannel | **F** | A1 + E |
| **C** | Beginning discourse segment | **N** | Literal modifier |
| **E** | Ending discourse segment | **S** | Stall/Filler |
| **B** | Back from a task | **?** | Cannot decide |
| **P** | Pivot: A1 + C | | |

Labelers were given examples of each category and labeled from both transcripts and speech. Inter-labeler reliability was measured by Fleiss' $\kappa$ [3] at 0.74 for the CARDS, 0.63 for the OBJECTS, and 0.69 for the whole corpus, where values between 0.6 and 0.8 correspond to substantial agreement. This study is based on MAJORITY labels, where at least 2 labelers assigned a token to the same class.

### 2.2. Feature extraction

To investigate the acoustic, prosodic and contextual characteristics of backchannels, we extracted features from the affirmative words themselves (uttered by the reference speaker), and from the preceding and following intonational phrases (ToBI level 4 phrases) uttered by the other speaker. (NB: 98%, N = 748 of our backchannels constitute a well-defined prosodic domain flanked by pauses.) We also labeled the position of the syllable boundary in 2-syllable words.

Continuous acoustic and durational features were extracted automatically using Praat ([2]) and then *z*-score normalized ($z = (X - mean) / st.dev.$), where *mean* and *st.dev.* were calculated from all speech uttered by the speaker in the session. We excluded the top and bottom 5% of the data in each domain to eliminate spurious pitch and energy data. In addition to minimum, maximum, mean of pitch and intensity, we extracted pitch slope, intensity slope, and stylized pitch slope, calculated over the whole phrase, its last 200 and 300ms, its second half, its last syllable, and the second half of its last syllable. The following ToBI labels were also extracted where available:

- last pitch accent, if any (e.g. H*, H+!H*, L*);
- break index (0–4);
- phrase accent and boundary tone, if any (e.g. L–L%, H–H%).

Contextual features included latency before and after the word, and duration of the interlocutor's preceding and following intonational phrases.

## 3. RESULTS

### 3.1. Lexical choice

Since different affirmative words can be used as backchannels, we wondered whether some items were chosen more frequently than others and whether this decision appeared to be speaker-dependent. Table 2 shows the frequencies of words labeled as backchannels (majority labels).

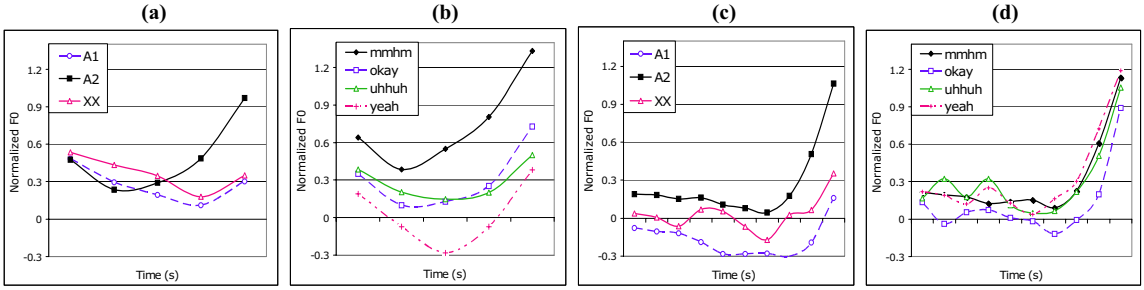
**Table 2:** Distribution of backchannels in percentages

| % | *mmhm* | *uhhuh* | *okay* | *yeah* | *right* | *alright* | *yep/yup* | *yes* | Σ |
|---|---|---|---|---|---|---|---|---|---|
| CARDS | 25.4 | 14.9 | 35.8 | 16.4 | 3.5 | 2 | 1 | 1 | 26.3 |
| OBJECTS | 62.5 | 20.1 | 8.7 | 6.9 | 1.2 | 0.4 | 0.2 | 0 | 73.7 |
| **Total** | 52.7 | 18.7 | 15.9 | 9.4 | 1.8 | 0.8 | 0.4 | 0.3 | 100 |

In the GAMES corpus, *mmhm* was the most common backchannel, followed by *uhhuh*, *okay* and *yeah*. All other words were used rarely. [5] reported that in their corpus, the *mmhm-uhhuh* category was the most frequent (46%), followed by *yeah* (27%) and *right* (9%). We also observe asymmetries between the CARDS and OBJECTS: *mmhm* and *uhhuh* are more common in OBJECTS than in CARDS, and together comprise over 80% of backchannels in OBJECTS. *Okay* and *yeah* show the opposite pattern. While both game types are oriented around a task, OBJECTS games appeared to engage participants in livelier interactions, which may explain this difference. Interestingly, [5] found *yeah* to be the most ambiguous word, conveying agreement, backchannel, incipient speakership, and yes-answer frequently. In our corpus, the most ambiguous word was *okay*.

There is some speaker variability in choice of lexical item for backchannels. While *mmhm* and *okay* are distributed fairly evenly among our speakers, only 5 of 13 speakers use *uhhuh* as a backchannel and almost half the tokens (43%) come from a single speaker. All tokens of *alright* and *right* also come from a single speaker.

**Figure 1:** F0 of **(a)** discourse markers, **(b)** backchannels: means of 5 equidistant intervals z-normalized. F0 of the last second of the intonational phrase preceding a **(c)** discourse marker, **(d)** backchannel: means of 100ms intervals z-norm.




### 3.2. Prosodic characteristics of backchannels

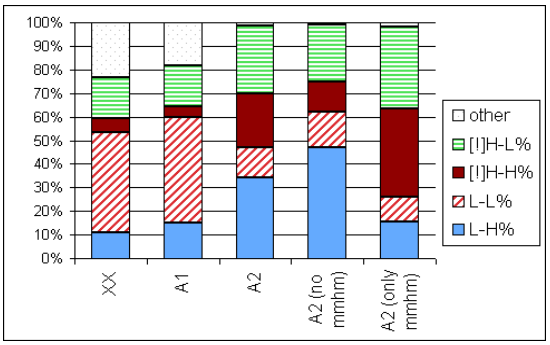
To determine whether backchannels were prosodically different from the other discourse functions of affirmative words, we collapsed some of our smaller categories, leaving only three groups: backchannels (A2, N = 763, 16%), agreements (A1, N = 2370, 50%), and XX (all other labels excluding literal modifiers, N = 1581, 34%). We tested for differences between the groups using multiple Kruskal-Wallis tests, and when a feature showed significant differences, we followed by Tukey pair-wise comparisons of mean ranks. Results showed (Table 3) that backchannels (A2) ***are*** strongly marked prosodically: they have higher pitch, intensity and pitch slope than both agreements (A1) and other functions (XX). Fig. 1a illustrates the differences in pitch contours. Backchannels are also longer than other functions (XX) and similar in duration to agreements (A1).

The prosodic difference between backchannels and other discourse functions of affirmative words is corroborated by examining ToBI labels. Chi-square tests showed that the intonational contour of backchannels is significantly different from other functions. All agreement words tend to be uttered with a H* pitch accent, but backchannels are ***more*** likely to have L+H* accents and ***less*** likely to have H+!H* and L* accents; $X^2$ = 89.4, df = 6, $p$ = 0. All affirmative words tend to have continuation rise (L-H%) or plateau (H-L%) endings, but backchannels are ***more*** likely to have a high boundary tone (H-H%) and much ***less*** likely to have a low tone (L-L%); $X^2$ = 262.5, df = 6, $p$ = 0. Note that *mmhms* are the most frequent backchannel in our corpus and tend to have H-H% endings, which may bias this result. When *mmhms* are excluded, the backchannels are best characterized by the ***in***frequency of L-L% and frequency of L-H%; $X^2$ = 266.7, df = 8, $p$ = 0. These results are summarized in Figure 2.

**Table 3:** Features best predicting discourse functions (Kruskal-Wallis: A1 vs. A2 vs. XX, $p$< 0.001; shaded cells show significant differences; Tukey, $p$ < 0.05).

| Attribute | median | | | A1-A2 | A1-XX | A2-XX |
|---|---|---|---|---|---|---|
| | A1 | A2 | XX | | | |
| duration.ms-z | -0.077 | -0.115 | -0.241 | | > | > |
| min.pitch-z | -0.608 | -0.103 | -0.488 | < | | > |
| mean.pitch-z | 0.043 | 0.347 | 0.181 | < | < | > |
| max.pitch-z | 0.792 | 1.065 | 1.011 | < | < | |
| min.int-z | -1.725 | -0.608 | -1.879 | < | > | > |
| mean.int-z | -0.626 | -0.133 | -0.652 | < | > | > |
| max.int-z | 0.026 | 0.243 | 0.054 | < | | > |
| pitch.slope | -60.1 | 39.6 | -63.8 | < | | > |
| pitch.slope.200ms | -53.0 | 167.5 | -86.9 | < | | > |
| pitch.slope.2ndHalf | -46.2 | 199.3 | -101.1 | < | > | > |
| sty.pitch.slope | -64.4 | 42.0 | -73.6 | < | | > |
| sty.pitch.slope.200ms | -50.1 | 59.6 | -73.7 | < | | > |
| sty.pitch.slope.2ndHalf | -46.4 | 57.0 | -73.1 | < | > | > |

**Figure 2:** Boundary tones of backchannels (ToBI)




We next compared the four most frequent backchannels among themselves (*mmhm*, *okay*, *uhhuh*, and *yeah*) and also *yeah* with the second syllable of *mmhm*, *okay*, *uhhuh* to see whether different lexical items used as backchannels are uttered similarly. We found the major difference to be in pitch (Fig. 1b): *mmhms* have higher pitch than the other three words, and *mmhm* is likely to have more rising pitch than *uhhuh* and *yeah*. In

addition, backchannel *okays* in our corpus have lower intensity, and *mmhm* and *okay* tend to have greater intensity slopes than *yeah*.

### 3.3. Context of backchannels

In our corpus, backchannels tend to follow intonational phrases (IPs) with rising pitch. In the ToBI labeled portion of the corpus, 73% of backchannels follow a rising intermediate or intonational boundary (H-, H-H%, L-H%). This trend is confirmed by looking at normalized F0 values in the last second of the IP preceding a backchannel (Fig. 1c). Backchannels (A2) are preceded by IPs that end in rising pitch starting with a low tone around 0.5s before the backchannel, presumably corresponding to a L- phrase boundary tone in ToBI.

In terms of predicting the lexical type of backchannel from the prosody of other speaker's preceding phrase, our data show that IPs with lower mean pitch and greater pitch slope tend to be followed by *okays* (Fig. 1d) and IPs with greater intensity tend to be followed by *uhhuhs*.

Durational features also correlate with discourse functions of affirmative words. Using the same tests as in Table 3 we found that latency to response is fastest for backchannels (A2) then agreements (A1) and then other (XX) ($p < 0.05$). Among backchannels, *okays* are produced with longer latency than *mmhm*, *uhhuh*, or *yeah* ($p < 0.05$). Also, the preceding IP's length and the backchanneling speaker's latency positively correlate with backchannel length ($r = 0.12$, $p < 0.001$, $r = 0.1$, $p < 0.01$).

Again looking at speaker variation, a one-sided *t*-test revealed that female speakers in our corpus tend to backchannel sooner when responding to female speakers than to males; $t = 3.06$, df = 91, $p < 0.005$. Females also resume speaking sooner after a backchannel from a female interlocutor than from a male; $t = -3.17$, df = 130, $p < 0.001$.

### 3.4. Effectiveness of backchannels

To measure how useful backchannels are in dialogue, we examined the correlation between the frequency of backchannels in each task of the OBJECTS games and the score subjects obtained in that task, as a rough objective measure of effectiveness. While there is no correlation between task score and task length (e.g., duration, number of speaker turns), the number of backchannels, normalized for task length, shows a weak positive correlation with task success; Pearson's $r \approx 0.14$, $p \approx 0.055$.

## 4. CONCLUSION

Our study of backchannels in task-oriented dialogue between Standard American English speakers shows that both prosodic and contextual factors distinguish backchannels from other affirmative words. Backchannels are generally higher in pitch and intensity with greater pitch slope than other affirmative words. While they are not different from agreements in duration, both differ from other uses of affirmative words in this feature. Also, backchannels in general tend to occur following a rising phrase from the interlocutor. These findings are strengthened by the fact that both discrete (ToBI) and gradient features provide complementary results.

Lexical choice for the backchannel appears somewhat speaker dependent, but may be predicted by the pitch and intensity of the interlocutor's preceding phrase. Backchannel behavior also appears to be influenced by the gender of both the speaker and hearer in the dialogue. And finally, subject performance on task was correlated with the number of backchannels the dyad produced, normalized for task length. Thus our findings suggest that not only are backchannels important to successful communication, but they show promise of being modeled effectively (recognized and produced naturally) in spoken dialogue systems.

## 5. ACKNOWLEDGEMENTS

This work was funded in part by NSF IIS-0307905. We thank Gregory Ward, Elisa Sneed, and Michael Mulley for help with collecting and labeling the data, and the anonymous reviewers for helpful comments and suggestions.

## 6. REFERENCES

[1] Beckman, M. E., Hirschberg, J. 1994. The ToBI annotation conventions. Ohio State University, 1994.

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

[3] Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378-382.

[4] Hockey, B. A. 1993. Prosody and the role of okay and uh-huh in discourse. *Proc. ESCOL*, 128-136.

[5] Jurafsky, D., Shriberg, E., Fox, B., Curl, T. 1998. Lexical, Prosodic, and Syntactic Cues for Dialog Acts. *Proc. ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers, 114-120.*

[6] Ward, N., Tsukahara W. 2000. Prosodic Features which Cue Back-channel Responses in English and Japanese. *J. of Pragmatics*, 32(8), 1177-1207.