A Regression-based Approach to Modeling Addressee Backchannels

Allison Terrell, Bilge Mutlu
Department of Computer Sciences, University of Wisconsin–Madison
1210 West Dayton Street, Madison, WI 53705, USA
{aterrell,bilge}@cs.wisc.edu

Abstract

During conversations, addressees produce conversational acts—verbal and nonverbal *backchannels*—that facilitate turn-taking, acknowledge speakership, and communicate common ground without disrupting the speaker's speech. These acts play a key role in achieving fluent conversations. Therefore, gaining a deeper understanding of how these acts interact with speaker behaviors in shaping conversations might offer key insights into the design of technologies such as computer-mediated communication systems and embodied conversational agents. In this paper, we explore how a regression-based approach might offer such insights into modeling predictive relationships between speaker behaviors and addressee backchannels in a storytelling scenario. Our results reveal speaker eye contact as a significant predictor of verbal, nonverbal, and bimodal backchannels and utterance boundaries as predictors of nonverbal and bimodal backchannels.

1 Introduction

Conversations involve a dynamic shifting of speakership, one party playing the role of the "speaker" and the other(s) the role of the "addressee" at any given moment (Goodwin, 1981; Levinson, 1988; Clark, 1996). In these roles, while speakers produce the majority of the conversational content, addressees play a major role in facilitating speakership by performing *backchannels*—verbal and nonverbal acts such as "uh huh" and head nods that indicate the addressee's understanding and involvement and acknowledge that the speaker has and may continue to have the floor (Yngve, 1970; Drummond and Hopper, 1993).

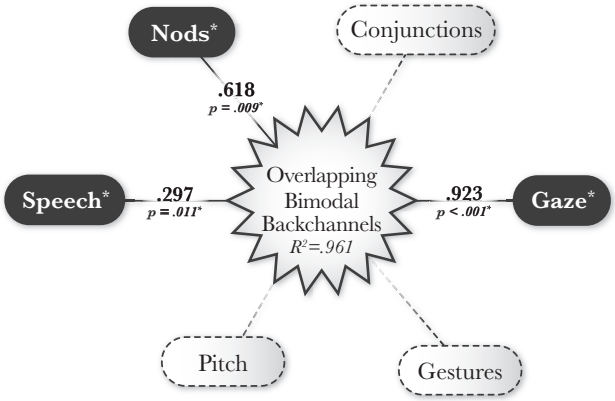

Figure 1: A mapping of the predictive relationships between speaker behaviors and overlapping bimodal addressee backchannels. β coefficients show the relative importance of significant predictors of backchannel behaviors.

Backchannels serve as a mechanism for cooperation between speakers and addressees to achieve efficient communication (Brunner, 1979; Grice, 1989) and to establish rapport (Drolet and Morris, 2000). The design of conversational technologies such as computer-mediated communication systems will have to facilitate the use of backchannel mechanisms to help their users achieve efficient conversations. Similarly, embodied conversational agents will have to use these mechanisms to achieve efficient interactions with their users. However, these developments require a deeper understanding of backchannel behavior and models of the relationship between backchannel acts and speaker behaviors.

Research across many communities including discourse processes, dialog systems, and human-computer interaction has explored the use of backchannels in conversations and sought to model the relationships between backchannel acts and other conversational processes using techniques that range from contingency analyses (Truong et al., 2011) to model training (Morency et al., 2010). In this paper, we propose a complementary, regression-based approach to untangle the *predictive* relationships between speaker behaviors and addressee backchannels. This approach provides us with an understanding of what speaker behaviors are significant predictors of addressee backchannels and of the relative contributions of each behavior in these predictions. The resulting models inform us of what speaker behaviors are important to support in interactive systems and communication technologies to facilitate addressee backchannels and complement finer-granulated analyses of specific backchannel mechanisms.

We contextualize our exploration in a storytelling scenario, which requires addressees to rely on and frequently use backchannels to participate in the discourse while maintaining consistency in conversational roles, using a multimodal data corpus collected from 24 dyads. Our analysis includes verbal and nonverbal backchannels, focusing on continuers and assessments in the verbal channel and head nods in the nonverbal channel. In the remainder of the paper, we review related work, describe our methodology, present our results, and discuss our findings and their implications for future research and the design of communication and interactive technologies.

2 Background

Conversations involve a cooperative process in which interlocutors manage the floor, negotiate turns, and provide feedback with the aid of subtle linguistic and extralinguistic cues—*backchannels*—that might not significantly contribute to the substance of the conversation (Yngve, 1970; Brunner, 1979; Grice, 1989; Drummond and Hopper, 1993). These backchannels allow parties, particularly addressees, to exchange information on their intentions and statuses and to participate in the conversation without disrupting ongoing speech (Morris and Desebrock, 1977; White, 1989). Backchannels differ from "backchannel inviting cues," which might indicate what might be an appropriate time for a backchannel (Gravano and Hirschberg, 2011). While backchannels are produced universally, individual characteristics such as gender (Helweg-Larsen et al., 2004) and cultural background (White, 1989; Ward and Tsukahara, 2000) significantly shape their production and interpretation.

2.1 Backchannel Cues

Researchers have sought to distinguish and categorize the wide range of backchannels based on how they are expressed by addressees (Jenkins and Parra, 2003) and how they contribute to the conversation(Young and Lee, 2004). The majority of research on backchannels considers *verbal* or linguistic cues and offers several categorizations. One of these categorizations distinguishes *continuers* from *assessments* (Young and Lee, 2004). Continuers are short, nondescript verbal segments such as "uh huh" and "yeah" that prompt the speaker to continue talking, while assessments are longer verbal segments such as "oh, wow" and "really?" that offer commentary or request clarification on the speaker's statements.

Another classification of verbal backchannels distinguishes among *non-lexical*, *phrasal*, and *substantive* backchannels (Iwasaki, 1997; Young and Lee, 2004). Non-lexical backchannels include vocalizations such as "hmm" or "uh huh" that offer little or no meaning but indicate the addressee's engagement in the conversation. Phrasal backchannels involve simple, well-established expressions such as "Really?" or "Are you serious?" that indicate acknowledgment. Finally, substantive backchannels involve the addressee taking the floor for brief periods and include repetitions, summary statements, clarifying questions about the speaker's speech, repair, and collaborative completions.

Research on backchannels also describes *nonverbal* or extralinguistic cues such as smiling (Brunner, 1979) and gaze (Rosenfeld and Hancks, 1980) as common backchannel behaviors that indicate agreement, understanding, or engagement in the conversation (Jenkins and Parra, 2003). Nodding is a particularly common nonverbal backchannel behavior that plays a range of roles from indicating agree-

ment to conveying sympathy and understanding with the speaker's perspective (Stivers, 2008). While verbal and nonverbal backchannels play similar communicative roles, the specific context of the conversation, such as whether the conversation involves a negotiation or a discussion, shapes how participants perform and interpret the two forms of backchannels (Jenkins and Parra, 2003). Addressees often display both verbal and nonverbal backchannels (Truong et al., 2011), such as concurrently nodding and saying "yeah" to express agreement.

2.2 Modeling Backchannels

Research on conversational backchannels involves a wide range of modeling approaches including *rule-based models* (Duncan, 1972), *contingency analysis* (Truong et al., 2011), and *trained models* (Morency et al., 2010) across a wide range of conversational contexts from telephone conversations (Ward and Tsukahara, 2000) to face-to-face interactions (Truong et al., 2011). Rule-based models capture relationships between backchannels and other conversational behaviors based on prototypical examples of commonly observed behaviors. Contingency analysis offers a quantitative basis for modeling these relationships through pairwise analyses of co-occurrences. Finally, statistical learning techniques allow researchers to train machine learning algorithms, such as Support Vector Machines (SVM) and Hidden Markov models (HMM), on data that capture these relationships in order to estimate the timing of backchannels.

2.3 Regression-based Modeling

While it remains unexplored in the context of modeling backchannel behaviors, regression-based approaches are commonly used in modeling complex relationships among many variables. In the context of modeling discourse and dialog, frameworks such as PARADISE (PARAdigm for DIalogue System Evaluation) build on regression-based approaches to identify predictive relationships between several elements of dialog and objective or subjective outcomes of the dialog (Walker et al., 1997). Researchers have used these frameworks to evaluate the effectiveness of spoken dialog in interactive systems (Foster et al., 2009; Peltason et al., 2012).

3 Method

Due to the broad range of verbal and nonverbal backchannels, we chose to focus on a limited subset of verbal and nonverbal cues, including continuers and assessments as verbal backchannels and head nods as nonverbal backchannels. Although there are numerous possible speaker behaviors, which may predict backchannels, we focused on six cues based on previous research: (1) speaker's *gaze* (directed toward the addressee), (2) *nods*, (3) *gestures*, (4) *speech* (whether the speaker is speaking or not), (5) *conjunctions* in the speaker's speech, and (6) *pitch variance* in the speaker's speech. These six predictors were then used to build models for five dependent variables: (1) *nonverbal backchannels*, (2) *verbal backchannels*, (3) *concurrent verbal and nonverbal backchannels* (e.g., a nod and an "OK" starting simultaneously), (4) *overlapping verbal and nonverbal backchannels* (e.g., a nod followed by an "OK" towards the end of the nod), and (5) *independent bimodal backchannels* (the presence of either verbal or nonverbal backchannels). We modeled the relationships between these predictors and dependent variables using stepwise regression.

3.1 Participants and Data Corpus

A total of 48 subjects from the University of Wisconsin–Madison participated in this study. They studied a diverse set of fields and were aged between 18 and 28. All participants were native English speakers. We assigned participants into dyads and conversational roles following a fully stratified design to control for the effects of gender composition of the dyads. We discarded data from one dyad, because the participants did not conform to the conversational roles that they were asked to follow. With this omission, our final dataset consisted of 23 dyads.

Our experimental setup followed common conventions of face-to-face conversations. Two participants unfamiliar with one another were seated across from each other at a "social distance" of five feet (Hall, 1963). An illustration of our experimental setup can be seen in Figure 2. The data collection equipment consisted of three high-definition video cameras at 1080p resolution and 30p frame rate, two high-fidelity lapel microphones, and an omni-

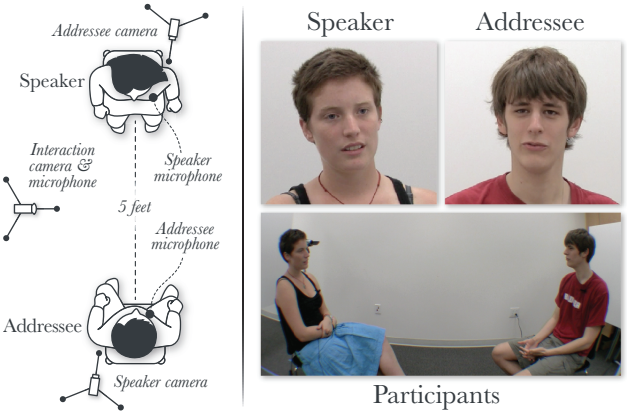

Figure 2: The experimental setup (left) shows the placement of the participants at a "social distance" and of the equipment for capturing data. The snapshots (right) show the vantage point from each of the three cameras.

directional microphone. Two of the video cameras were positioned across from each participant, capturing their upper torso from a direct frontal angle, while the lapel microphones captured their speech. The third camera and the omni-directional microphone recorded the speech and nonverbal behaviors of both participants from a side angle. The final datas corpus consisted of 1 hour and 31 minutes of audio and video. The average video length was 3 minutes and 57 seconds.

3.2 Procedure

The experimental task involved partaking in a *storytelling* scenario that aimed to elicit a wide range of behavioral and interactional mechanisms. In this scenario, one of the participants took on the role of the speaker and narrated the plot of their favorite movie to the second participant who took on the role of the addressee. We expected this scenario to provide us with a rich context to observe backchannels.

Participants were first given a brief description of the experiment and asked to review and sign a consent form. The experimenter then seated the participants, assigned them conversational roles, and set up the data collection equipment. Participants first performed an acclimation task (getting to know one another) that was not considered part of the experimental task. The participants then performed the storytelling scenario. Following the experiment, the experimenter debriefed the participants. Participants were paid $10 for their time.

3.3 Measurements

Based on a preliminary analysis of our data, we identified five forms of addressee backchannels as dependent variables: (1) nonverbal backchannels, (2) verbal backchannels, (3) concurrent verbal and nonverbal backchannels (e.g., a nod and an "OK" starting simultaneously), (4) overlapping verbal and nonverbal backchannels (e.g., a nod followed by an "OK" towards the end of the nod), and (5) independent bimodal backchannels (either verbal or nonverbal backchannels).

Our independent variables consisted of speaker behaviors that previous research suggested as likely predictors of addressee backchannels and that a real-time interactive system might be able to capture and interpret. These variables included visible and audible features from the speaker's movements and speech, such as the presence or absence of speech and pitch variability, and specific linguistic features that might signal discourse structure, such as conjunctions. Drawing on these considerations, our analysis included speaker's *gaze* (directed toward the addressee), *nods*, *gestures*, *speech* (whether the speaker is speaking or not), *conjunctions* in the speaker's speech, and *pitch variance* measurements of the speaker's speech.

In our measurement of pitch, we sought to capture computationally feasible, high-level intonational characteristics of the speech by calculating the variability in pitch in the entire conversation. Low pitch variability indicated more monotonous speakers, whereas high pitch variability represented more expressive speech. This measure was calculated by finding the average pitch of the speaker throughout the conversation and aggregating the difference between the average pitch and the pitch value at each frame, as expressed below:

$$pitch\_variance = \sum_{i=0}^{n} |\overline{pitch} - pitch_i|$$

Here, the number of measurements in the conversation is represented by n; each individual measurement is represented by i; the speaker's average pitch in the entire conversation is represented by $\overline{pitch}$; and the pitch value at each individual measurement is represented by $pitch_i$.

The data was labeled using a combination of manual and computational techniques. All speaker and

Measure (y)	*Function* ($\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + e$)	R^2	*Significance*
Nonverbal backchannels	$.138 + .635 \times \mathcal{N}(gaze) + .374 \times \mathcal{N}(speech) + .089$	.911	*gaze* $p < .001$ *speech* $p = .008$
Verbal backchannels	$.034 + .875 \times \mathcal{N}(gaze) + .067$	.977	*gaze* $p < .001$
Concurrent bimodal backchannels	$.019 + .471 \times \mathcal{N}(gaze) + .822 \times \mathcal{N}(speech) + .059$	.940	*gaze* $p < .001$ *speech* $p < .001$
Overlapping bimodal backchannels	$.013 + .923 \times \mathcal{N}(gaze) + .618 \times \mathcal{N}(nods) + .297 \times \mathcal{N}(speech) + .061$	.966	*gaze* $p < .001$ *nods* $p = .009$ *speech* $p = .011$
Independent bimodal backchannels	$.134 + .483 \times \mathcal{N}(gaze) + .212 \times \mathcal{N}(pitch) + .074$	.896	*gaze* $p < .001$ *pitch* $p = .014$

Figure 3: The final models for each dependent variable after elimination in the stepwise regression analysis including only the significant predictors. Gaze was a significant predictor in all five models. Speech was significant in three models. Pitch variability and nods each significantly predicted one type of backchannel.

addressee utterances were transcribed using Praat. Speech and conjunctions measurements were drawn from this transcription. Only pauses that were longer than 500 milliseconds were considered as absence of speech; speech segments that were separated by shorter pauses were combined into a single segment. The pitch variability was automatically extracted using Praat. A primary coder labeled 100% of the remaining attributes (addressee nods, speaker nods, speaker gestures, and speaker gaze). To evaluate reliability, a second coder labeled 10% of a randomly sampled subset of the data. The inter-rater reliability showed substantial agreement for all attributes; addressee nods (94% agreement, Cohen's $\kappa = 0.72$), speaker nods (92% agreement, Cohen's $\kappa = 0.71$), speaker gesture (87% agreement, Cohen's $\kappa = 0.67$), and speaker gaze (96% agreement, Cohen's $\kappa = 0.75$).

All variables except pitch variability were binary: 0 for *not occurring* and 1 for *occurring* of events. Pitch variability was a normalized continuous variable that varied between 0 and 1. We considered variables as co-occurring when they overlapped with each other within a window that spanned 200 milliseconds before the onset and after the end of each variable, following criteria from previous research (Truong et al., 2011). The data corpus included measurements of all variables every 33.3 milliseconds.

The data corpus for each dependent variable included aggregate counts of measurements for all variables for each video. The aggregate counts for each video were normalized by dividing them by the length of the video in seconds. Finally, each variable across all videos were normalized to vary between 0 (least frequent) and 1 (most frequent). The resulting data corpus included five data tables of size $23x7$ (data from 23 dyads on seven variables—the dependent variable and six predictors) for five types of backchannel behaviors.

3.4 Analysis

Our analysis followed a *stepwise multiple linear regression* to model the relationships between our predictors and dependent variables. Each analysis started with the following linear form:

$$y = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n) + e$$

Here, β_0 is a constant, whereas $\beta_1 \ldots \beta_n$ are coefficient weights for each of n predictors. The values of each predictor for each measurement are represented by $x_1 \ldots x_n$. The error term for the model is e, which is assumed to be mean zero and independent and identically distributed (i.i.d.).

Our use of stepwise regression followed a *backward elimination* algorithm in which the final model is constructed by gradually excluding predictors that do not sufficiently contribute to the model. For the purposes of our study, we excluded any predictor with a p-value above .25. The final model is comprised of predictors left which are statistically sig-

	Nonverbal backchannels ($R^2 = .912$)		*Verbal backchannels* ($R^2 = .968$)		*Concurrent bimodal backchannels* ($R^2 = .938$)		*Overlapping bimodal backchannels* ($R^2 = .962$)		*Independent bimodal backchannels* ($R^2 = .889$)	
Predictor	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
Nods	.297	.542	.209	.242	.332	.227	**.579**	**.015**	.530	.205
Gaze	**.621**	**< .001**	**.952**	**< .001**	**.574**	**.001**	**.848**	**< .001**	**.542**	**.001**
Gestures	−.134	.275	−.033	.834	.052	.735	.235	.113	−.066	.489
Speech	**.388**	**.043**	.067	.644	**.631**	**.005**	.205	.091	.072	.775
Conjunctions	.129	.779	.323	.372	n/a	n/a	−.191	.622	.397	.325
Pitch	.012	.148	.003	.721	−.001	.761	.009	.655	.402	.118

Figure 4: The results of the model for each dependent variable before elimination in the stepwise regression analysis.

nificant ($p < .050$). The β coefficients in the model provide the relative contribution of each independent variable in predicting the dependent variable. Our analysis considered the number of addressee backchannels that occurred in each dyad as the metric of success.

3.5 Results

In all five of our models, the independent variables accounted for a significant proportion of variance in the dependent variables, varying between 89.9% and 96.6%. These results are summarized in Figure 3.

In the first model, speaker behaviors accounted for a significant portion of addressee nonverbal backchannels, $R^2 = .911, F(2,20) = 113.6, p < .001$. Speaker gaze and speech significantly predicted these backchannels, $\beta = .635$, $t(21) = 6.02$, $p < .001$ and $\beta = .374$, $t(21) = 2.90$, $p = .008$, respectively. Gaze also significantly predicted addressee verbal backchannels, $\beta = .875$, $t(22) = 27.24$, $p < .001$, and explained a significant portion of the variance in them, $R^2 = .977, F(1,21) = 702.5, p < .001$.

Results from the third model showed that gaze and speech explained a significant proportion of the variance in concurrent bimodal backchannels, $R^2 = .940, F(2,20) = 172.3, p < .001$, and significantly predicted these backchannels, $\beta = .471$, $t(21) = 3.98$, $p < .001$ and $\beta = .822$, $t(21) = 7.92$, $p < .001$, respectively. In the fourth model, speaker behaviors explained a significant proportion of the variance in overlapping bimodal backchannels, $R^2 = .966, F(3,19) = 180, p < .001$. Speaker gaze, speech, and nods were significant predictors of these backchannels, $\beta = .923$, $t(20) = 12.3$, $p < .001$, $\beta = .297$, $t(20) = 2.80$, $p = .011$, $\beta = .618$, $t(20) = 2.93$, $p = .009$, respectively.

Finally, results from the fifth model showed that speaker behaviors explained a significant proportion of the variance in independent bimodal backchannels, $R^2 = .896, F(2,20) = 94.63, p < .001$. The speaker's gaze and the variability in the pitch of the speaker's speech significantly predicted these addressee behaviors, $\beta = .483$, $t(21) = 6.74$, $p < .001$ and $\beta = .212$, $t(21) = 2.83$, $p = .014$, respectively.

4 Discussion

The results of our statistical analysis show key relationships between speaker behaviors and addressee backchannels, reaffirming findings from previous studies and revealing new relationships. The paragraphs below provide a discussion of these findings and support them with examples of addressee backchannels that we frequently observed in our data. These examples are illustrated in Appendix A in three episodes of interaction. We also discuss the implications of our approach for modeling conversational mechanisms.

Our results are summarized in Figures 3 and 4, which show our final models after elimination and the models before elimination, respectively. The results in Figure 3, consistent with previous work (Bavelas et al., 2002), highlight the importance of gaze in eliciting addressee backchannels. Gaze is

included in all five of our models and is consistently the most important predictor of addressee backchannels in four of our five models. In Appendix A, all six instances of the addressee backchannels across three illustrated episodes occur either when the speaker is looking toward the addressee or almost concurrently with the speaker shifting gaze away from the addressee.

The results also show speech to be a significant predictor of addressee backchannels. Three of our models included speech as a predictor, which suggests that more frequent pauses in speech provides the addressee with more opportunities to provide backchannels; that frequent pauses prompt addressees to provide more backchannels to facilitate the continuation of the speaker's speech; and/or that the addressees produce more backchannels, because speakers present more information. Four instances of backchannels shown in Appendix A occur immediately after an utterance has ended, which exemplify pauses as opportune moments for the addressee to produce backchannels.

The significance of pitch variability in predicting independent bimodal backchannels offers a different perspective on the relationship between attributes of speaker pitch and addressee backchannels than previous research does. Although previous work suggested that pitch attributes do not have a significant relationship with addressee backchannels in face-to-face conversations (Truong et al., 2011), pitch variability significantly predicted independent bimodal backchannels in our models. We speculate that pitch variability captures the speaker's overall ability to engage their addressees in their speech and, thus, predicts addressee backchannels. However, our results show that this predictive relationship only exists with independent bimodal backchannels and not with verbal or nonverbal backchannels. This discrepancy might be a result of variability across individuals in their preferences to use verbal and nonverbal backchannels, which is not captured by our models for these individual backchannels but is captured by the model that considers either type of backchannels.

Speech did not significantly predict the addressee's verbal or independent bimodal backchannels, while it predicted nonverbal and concurrent and overlapping backchannels. This finding suggests that frequent pauses in speech elicit primarily nonverbal backchannels and elicit verbal backchannels only in the presence of nonverbal backchannels. A possible explanation of this finding is that addressees might prefer nonverbal backchannels to verbal backchannels when they wish to facilitate the continuation of speech.

A key contribution of our work is an exploration of the relationship between verbal and nonverbal backchannels by modeling the concurrent onsets and overlaps between these backchannels. These models indicate that gaze and speech are significant predictors of concurrent onsets and overlaps in verbal and nonverbal backchannels and that speaker nods also significantly predict overlaps.

Our analysis also identified overlapping bimodal backchannels as a new form of backchannel behavior that has not been considered by previous research (Truong et al., 2011). These backchannels involve the addressee producing a nonverbal backchannel towards the end of the speaker's speech and then producing a verbal backchannel when the speaker had stopped talking. We speculate that this behavior allows the addressee to express agreement during the speaker's speech using nonverbal backchannels without disrupting the speech and reassert agreement using verbal backchannels when the speaker's utterance is completed. Episode B in Appendix A illustrates an instance of overlapping bimodal backchannels.

A final contribution of this work is an illustration of the use of a regression-based approach in modeling predictive relationships between speaker behaviors and addressee backchannels. This approach allowed us to explore the relationships among many aspects of speaker and addressee behavior and to quantify the relative significance of each aspect of the speaker's behaviors in predicting addressee behaviors. Our results confirmed findings from previous research and produced new findings, revealing novel relationships between these behaviors. These relationships will serve as a basis for future research to create more nuanced models of speaker and addressee behavior. They will also inform the design of future communication technologies and interactive systems that incorporate mechanisms to support the communication of key predictors of addressee backchannels.

While the primary goal of our study was to better understand relationships among conversational behaviors, our models might also serve as coarse estimation models. The models shown in Figure 3 might be used to estimate $\hat{y}$—how frequently addressee backchannels should appear—using the predictor coefficients β and values for known speaker behaviors x. These estimations might complement finer-granulated models of backchannel mechanisms in generating opportune backchannel behaviors for artificial agents and predict when these backchannels might occur in human-computer interaction scenarios.

4.1 Limitations

Our work also has a number of limitations. First, because our approach uses aggregate counts of behaviors from the entire interaction, it does not account for the temporal relationships among these variables. Therefore, the insights offered by our approach are limited to high-level conclusions on the relationships between these behaviors and illustrations of these relationships in example episodes of interaction. Future work should include complementary modeling techniques to build finer-granulated models of backchannel mechanisms.

Although participants in each conversation were explicitly assigned to one of the roles of speaker and addressee, we did not specifically tell addressees not to speak, which led to a greater amount of variability in their participation in the conversation, some offering up their opinions or asking questions throughout the speaker's story and others limiting their behaviors to a small number of backchannels. While this variability enabled more natural conversations, this lack of control might have limited the power of our statistical models.

In this paper, we focused on a set of high-level predictors that allow for real-time capture and interpretation, ignoring underlying conversational mechanisms such as repair, which might also serve as significant predictors of backchannels. The relationships between these mechanisms and backchannel behavior would be a fruitful area of exploration for future research.

Finally, the generalizability of our results suffers from the limited extent of the conversational context and participation structure of our experimental setup. Future work should seek to extend this exploration to a broader set of conversational settings, such as interview and discussion scenarios, and participation structures, such as multi-party conversations.

5 Conclusion

Backchannels are essential behaviors for achieving fluent and effective conversations. Gaining a deeper understanding of how these behaviors shape conversations might offer key insights into the design of technologies such as computer-mediated communication systems and embodied conversational agents. In an exploratory study, we used a stepwise regression approach to model the relationships between various types of addressee backchannels and speaker behaviors in a storytelling scenario. We found that gaze significantly predicted all types of backchannel behaviors including verbal, nonverbal, and bimodal backchannels. Our results also showed that speech, speaker nods, and pitch variability predicted some types of backchannel behaviors. While these results have some limitations due to our methodological choices, they suggest directions for future work and offer preliminary insights toward a deeper understanding of backchannel behaviors and how interactive systems and communication technologies might be designed to support these behaviors.

Acknowledgments

We would like to thank Faisal Khan for his help in data collection and processing. This work was supported by National Science Foundation award 1149970.

References

J.B. Bavelas, L. Coates, and T. Johnson. 2002. Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580.

L.J. Brunner. 1979. Smiles can be back channels. *Journal of Personality and Social Psychology*, 37(5):728.

H.H. Clark. 1996. *Using language*. Cambridge University Press.

A.L. Drolet and M.W. Morris. 2000. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50.

K. Drummond and R. Hopper. 1993. Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on Language and Social Interaction*, 26(2):157–177.

S. Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

M.E. Foster, M. Giuliani, and A. Knoll. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proc ACL/AFNLP*, volume 2, pages 879–887. Association for Computational Linguistics.

C. Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press New York.

A. Gravano and J. Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

P. Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

E.T. Hall. 1963. A system for the notation of proxemic behavior. *American anthropologist*, 65(5):1003–1026.

M. Helweg-Larsen, S.J. Cunningham, A. Carrico, and A.M. Pergram. 2004. To nod or not to nod: An observational study of nonverbal communication and status in female and male college students. *Psychology of Women Quarterly*, 28(4):358–361.

S. Iwasaki. 1997. The northridge earthquake conversations: The floor structure and the 'loop' sequence in japanese conversation. *Journal of Pragmatics*, 28(6):661–693.

S. Jenkins and I. Parra. 2003. Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal*, 87(1):90–107.

S.C. Levinson, 1988. *Putting linguistics on a proper footing: Explorations in Goffman's concepts of participation.*, pages 161–227. Oxford, England: Polity Press.

L.P. Morency, I. de Kok, and J. Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84.

D. Morris and G. Desebrock. 1977. *Manwatching: A field guide to human behaviour*. HN Abrams New York, NY.

J. Peltason, N. Riether, B. Wrede, and I. Lütkebohle. 2012. Talking with robots about objects: a system-level evaluation in hri. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, HRI '12, pages 479–486, New York, NY, USA. ACM.

H.M. Rosenfeld and M. Hancks, 1980. *The nonverbal context of verbal listener responses*, pages 193–206. The hague: Mouton Publishers.

T. Stivers. 2008. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction*, 41(1):31–57.

K.P. Truong, R.W. Poppe, I.A. de Kok, and D.K.J. Heylen. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proc Interspeech*, pages 2973–2976. International Speech Communication Association.

M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proc EACL*, pages 271–280. Association for Computational Linguistics.

N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207.

S. White. 1989. Backchannels across cultures: A study of americans and japanese. *Language in society*, 18(1):59–76.

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

R.F. Young and J. Lee. 2004. Identifying units in interaction: Reactive tokens in korean and english conversations. *Journal of Sociolinguistics*, 8(3):380–407.

Appendix A. Contextual Examples

Below are three example episodes drawn from our data. Each episode displays all occurrences of all the predictors we measured in real time. All six instances of backchannels highlight the importance of the speaker's gaze and speech in eliciting addressee backchannels.