# Automatically Generated Prosodic Cues to Lexically Ambiguous Dialog Acts in Multiparty Meetings

**Sonali Bhagat**[*] **Hannah Carvey**[*] **Elizabeth Shriberg**[*†]

[*] International Computer Science Institute, Berkeley, CA, USA
[†] SRI International, Menlo Park, CA, USA
{*sonalivb,hmcarvey,ees*}*@icsi.berkeley.edu*

## ABSTRACT

We investigate whether automatically extracted prosodic features can serve as cues to dialog acts (DAs) in naturally-occurring meetings. We focus on the classification of four short DAs, all of which can be conveyed by the same words. DAs were hand-labeled based on the discourse context. Results for classifiers trained on automatically extracted prosodic features show significant associations with DAs in unseen test data. Furthermore, the specific features used depend on the classification task at hand. Results shed light on the relationship between discourse function and prosody, and could be used to aid automatic processing for natural dialog understanding.

## 1 INTRODUCTION

One way to advance the automatic understanding of spontaneous dialog is to classify utterances according to their dialog act (DA) [2, 4, 5, 6, 8, 9, 10]. A DA refers to the role of a specific utterance in a conversation; for example, an utterance could be a statement, question, or command. Information about DAs could be used to aid the automatic summarization and browsing of conversation.

Previous work on automatic DA classification has typically been limited to two-party conversations or to restricted domains. In this paper we investigate cues to DAs in naturally-occurring meetings. Meetings offer an especially interesting context for such analyses because they are rich in diverse DAs, and display complex discourse structure. As a first investigation of DA modeling for this corpus, we focus specifically on four short DAs that can each be expressed by the same words, but that occur in quite different discourse contexts:

*Acceptance/Agreement* **(AGR)**: A DA that expresses acceptance of, or agreement with, the propositional content of another speaker's prior suggestion, statement, or question. For example:

Ann: Let's skip next week's meeting.
Bob: *Yeah*.
Eve: *Okay*.

*Acknowledgment* **(ACK)**. A DA that provides a confirmation of understanding of another speaker's utterance, rather than expressing agreement on the utterance's propositional content itself. E.g., in the following example, switch location is not something Bob can agree or not agree with:

Bob: Where's the switch? *Okay*.
Eve: Here, at the back.

*Backchannel* **(BAC)**. A **BAC** does not agree with or acknowledge specific information in the current talker's speech, but rather signals attention by the listener (similar to a nod of the head). Thus it does not take the floor but merely serves to encourage the current talker to continue:

Bob: We ran two jobs. But both are really slow.
Dave: *Yeah*. *Uhhuh*.

*Floor-grabber* **(FG)**. **FG**, like **BAC**, does not indicate explicit agreement with or acknowledgment of, prior specific content. **FG**s differ from **BAC**s, however, in that **FG**s function to grab the floor rather than encouraging the current talker to keep it.

Bob: I got a win with that, so I'm happy.
Dave: *Yeah*, but did you normalize?

In all cases above, note that the DA is defined by the prior and following DA context—not by differences in prosody *per se*. Also note that for the first two cases, the speaker has the floor when uttering the DA (there is an expectation of a response at that point in time), whereas in the second two cases, the speaker does not have the floor when uttering the DA.

Each of these four DAs can be expressed with any one of the following words in our corpus: *"yeah"*, *"right"*, *"okay"*, and *"uhhuh"*, although the distribution of words over DAs is not uniform, see Table 1.) We focused on this small set of words because they cover a fair amount of the data without introducing DAs having quite different characteristics. For example, other DAs used as **F**s include filled pauses and false starts.

We ask whether there are inherent, automatically ex-

**Table 1:** Frequency of Four DAs by Word. **AGR** = acceptance, **ACK** = acknowledgment, **BAC** = backchannel, **FG** = floor-grabber. % Tot = percentage of total cases in 20-meeting corpus covered by the 4 DAs / 4 words.

| Word | AGR | ACK | BAC | FG | % Tot. |
|---|---|---|---|---|---|
| *yeah* | 440 | 90 | 1759 | 184 | 43.7 |
| *right* | 175 | 112 | 335 | 40 | 42.8 |
| *okay* | 13 | 411 | 212 | 54 | 42.4 |
| *uhhuh* | 28 | 19 | 873 | 4 | 43.5 |
| **% Tot.** | 33.2 | 34.1 | 47.0 | 22.7 | |

tractable prosodic differences among the four DAs, and what types of prosodic features characterize the various distinctions. We explore how well classifiers perform on an all-way task as well as on various meaningful subtasks. We present results for experiments in which a range of prosodic feature types is available to the classifier, as well as experiments constrained to use only one of the feature types.

## 2 METHOD

**Speech data**. Our data come from the ICSI Meeting Recorder corpus [7]. The corpus contains the audio files and word-level transcripts of 75 meetings. Each meeting is 30 to 80 minutes in duration and has between 3 and 10 participants. Speakers are recorded using both close-talking and far-field microphones; we use the former for this work. We use data from 20 meetings that have been completely annotated for DAs according to [1]. Frequencies of the DAs and words of interest here are listed in Table 1. We used 16 meetings for training and 4 for testing, with no meeting overlap (although there is speaker overlap).

**DA annotation**. Meetings were hand-annotated by four labelers after development of a suitable annotation scheme [1] comprising 58 different tags. This scheme is an adaptation of the SWBD-DAMSL Coding Manual for one-on-one telephone conversations [8], extended to cover the DAs found in the meeting corpus. Annotators hand-labeled time-aligned transcripts while listening to the sound files. As noted in the introduction, it is important to point out that the prosody of the four DAs studied here was *not* the sole determinant of DA class. Rather, the DAs of prior and following context determined the class of these short DAs. For example, in the excerpt demonstrating class **AGR** in the Introduction, Bob's and Eve's responses to Ann's suggestion cannot be labeled as **BAC**s, because **BAC** does not make sense as a response to a suggestion. Prosody is however often used in determining the DA class of surrounding context.

All utterances (not only those studied here) were labeled for DAs at this stage. Interlabeler reliability for the DAs used in this study was computed as Kappa (the amount of agreement after adjusting for chance agreement). We obtained a Kappa of .74, which is quite good for this type of task [3].

**Features.** For each DA, we extracted a vector of prosodic features, based on time alignments from an automatic speech recognition system run in forced aligment (true words) mode. We used true words in order to estimate the inherent contribution of prosodic features in the best case scenario, i.e. without confounding effects from faulty word alignments or segmentations due to word recognition errors.

*Duration features* included the phone-normalized duration of the longest normalized vowel in the word. Normalization included ratio and Z-scores based on the duration of the phone over all words.

*Energy features* included the maximum, minimum, and mean energy in the word, after straight-line approximations of energy contours, both with and without voiceless regions (based on voicing estimates from a pitch tracker). Also included were features capturing the shape of energy contour over the course of the word (rising or falling). We note that these features were not normalized for the speaker/channel (although they should be), as it is not clear how to normalize over all speech when different speakers produce different ratios of the various DAs.

*Pitch features* included the maximum, minimum, mean, first, and last F0 value in the word, after stylization to remove halving and doubling errors, normalized by a speaker "baseline F0" value obtained from a lognormal tied-mixture model of pitch [11]. These features capture how high or low a word is within a speaker's pitch range. Also included were features capturing the shape of the pitch contour over the word (rising or falling).

*Pause features* included the time between the beginning or end of the DA in question, and the end or start time of the preceding or following DA, respectively. Because our DAs differ in terms of turn-taking and floor-grabbing, we looked at both the pauses with respect to the *same* speaker's DAs (SS), and those with respect to a *different* speaker's DAs (DS). Thus, each DA is associated with four different pause values (pre-DA, post-DA, crossed with SS, DS).

**Classifiers**. We trained decision trees as classifiers to yield posterior probabilities of the DAs given the prosodic features. Decision trees offer the advantage that they can be inspected to gain an understanding of how features behave. To help overcome the problem of greediness, we wrapped a feature subset selection algorithm around the standard tree growing algorithm, as in [11]. To make the trees sensitive to prosodic features in the case of highly skewed class sizes, we trained on a resampled version of the target distribution in which all classes have equal prior probabilities. This also allows for direct comparison of results across tasks.

## 3 RESULTS AND DISCUSSION

We examined five different tasks, including a 4-way classification of DAs using all words, a 4-way classification of DAs using only the word "yeah", and three 2-way classifi-

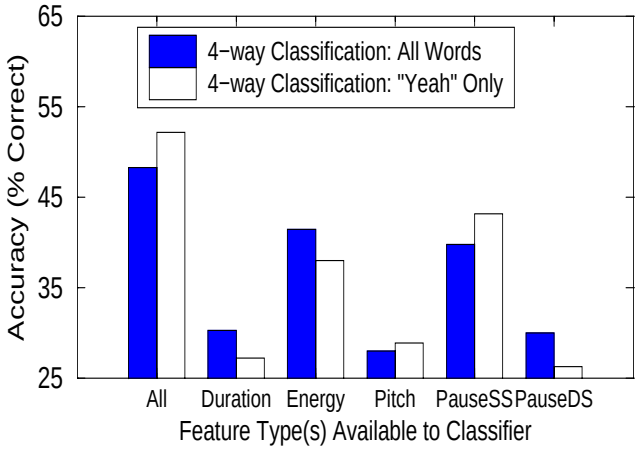


**Figure 1:** Results for all-way classification: Agreements vs. Acknowledgments vs. Backchannels vs. Floorgrabbers, using equal class priors. Chance performance = 25%.

cations using all words. For each of the five tasks we ran six different classifiers: one classifier with all feature types available, and five additional classifiers each with only one feature type available.

**All-way classification**. Model accuracies for the all-way classification are shown in Figure 1. Note that because class priors are equated, chance is 25% for this task. In addition to the main task, which collapsed over different words, we ran a second task for the case of "yeah" only. As shown in 4 Table 1, "yeah" has a reasonable number of tokens in each class. By comparing DAs in this second task, we know that differences in any prosodic features across DAs cannot be attributed to differences in the distribution of specific words.

A first observation from Figure 1 is that overall results are similar for all words and for "yeah" alone, despite the much smaller amount of data in the latter condition. This suggests that the general prosodic patterns captured are not dependent on the words. Examination of confusion matrices for both experiments reveals that the DA with highest accuracy is **FG**. Second, the best two feature types for this task are energy and pauseSS. Although it is difficult to infer the contribution of particular features to particular distinctions in this four-way task, it appears from inspection of the classifers that **FG**s are distinguished from all other DAs by high energy and a short following same-speaker pause (the latter expected given their function).

**Two-way classifications**. To better understand which features aid which DA distinctions, we ran three two-way classifiers on specific DA pairs that contrast on a dimension of interest. One such contrast is **AGR** versus **ACK**. These DAs differ with respect to their reference to propositional content (**AGR** refers to agreement with content, while **ACK** does not). A second such contrast is **ACK** versus **BAC**. In this case, the contrast is based on a difference in what the speaker is acknowledging (confirmation of an answer in the case of **ACK**, but mere verbal "head-nodding" of another's ongoing talk in the case of **BAC**). Finally, we were interested in the comparison between **BAC** and **FG**, since they function in opposite ways (**BAC** to encourage another speaker to continue; **FG** to grab the floor). Figure 2 shows results for the three two-way classifications, in all cases collapsed over words.

Figure 2 clearly shows that the contribution of specific prosodic feature types depends on the distinction at hand. Looking at the first task (**AGR** versus **ACK**) only, we see that energy and pitch are both good features when used alone. Interestingly, the all-features tree for this task, which has a much higher accuracy than either energy or pitch, uses pauseSS features in addition. The top feature usages for that tree are: pitch 47.1%, pauseSS 32.8%, and energy 10.9%. The **AGR**s are utttered with higher pitch, are followed by shorter same-speaker pauses, and are associated with more falling energy contours than are **ACK**s. This is consistent with a more emphatic rendering of **AGR**s than of **ACK**s, and a greater probability of continuation by the same talker after **AGR**s than after **ACK**s.

For the second task, **BAC** versus **ACK**, duration, energy, and pauseSS are the most successful features. The result for the all-features condition (69.4%) is only slightly better than that for the duration only condition (66.7%); the better result is achieved by a combination of pause, duration, and pitch features, in that order of usage. Results show that **BAC**s are followed by longer same-speaker pauses than are **ACK**s. Pitch features are higher for **ACK**s than **BAC**s, consistent with the greater content in the former. Duration is longer for **BAC**s than for **ACK**s, a result that was unexpected given that longer durations are usually associated with greater content in most prosodic phenomena (e.g., pitch accents). Perhaps it is the case that **BAC**s are drawn out (just as a head nod may be performed slowly) to indicate ongoing attention.

Finally, we turn to results for the **BAC** versus **FG** task. Accuracy is notably high for this condition, particularly if using only pause information. This is somewhat expected, since by definition, **FG**s attempt to grab the floor, and therefore are much more likely to be followed by speech from the same talker. What is interesting about these results is that **FG**s differ from **BAC**s in other features as well, including energy, pitch, and duration. Thus, **FG**s are not **BAC**s that turn into **FG**s; rather they are more like **FG**s parading as **BAC**s to "soften" an interruption. **FG**s are much higher in energy and pitch than are **BAC**s, which is consistent with their function, since they need to be noticed in order to be useful in obtaining the floor. Also noteworthy is that although **FGs** are louder and higher in pitch than **BAC**s, they are *shorter* than **Bs** in duration. As noted above, typically one finds that duration correlates positively with other prosodic features. Here, speakers do not draw out their **FG**s as they do their **BAC**s. This suggests that in producing **FG**s, speakers may be to trying to minimize the time as-

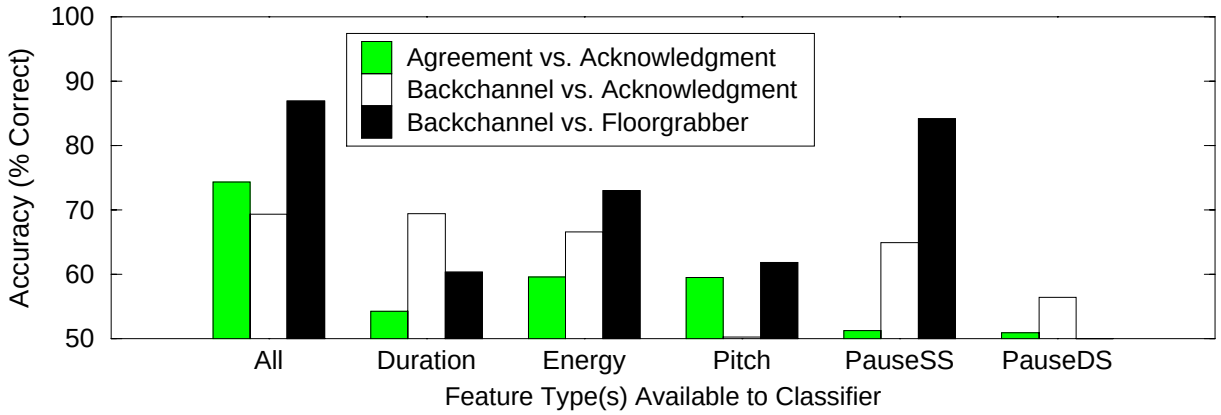


**Figure 2:** Results for two-way classifications, collapsing over words. Class priors are equated, thus chance = 50%.

sociated with an obvious interruption, while nevertheless making sure that the interruption is heard.

## 4 CONCLUSION

We have shown that automatically extracted prosodic features provide cues that distinguish among four types of lexically ambiguous DAs in naturally-occurring multiparty meetings. While best performance is achieved by combining different types of prosodic features, the performance of classifiers using only particular features reveals that different DA distinctions are cued by different features. From a theoretical perspective, analyses of specific feature patterns by task reveal interesting ways in which speakers use prosody to convey different utterance functions in natural conversation. From the applied side, since the DAs examined correlate with surrounding DA context, these results suggest that prosody can provide useful information for the automatic modeling of meeting data.

## ACKNOWLEDGMENTS

We thank Chuck Wooters for DA annotation software, Luciana Ferrer for prosodic feature extraction, and Raj Dhillon and Ashley Krupski for DA annotation. This work was supported by an ICSI/SRI/Columbia/UW NSF ITR, an ICSI DARPA Communicator project, SRI NSF Award IRI-9619921 and SRI NASA Award NCC2-1256. The views herein are those of the authors and do not reflect the views of the funding agencies.

## REFERENCES

[1] S. Bhagat, H. Carvey, R. Dhillon, A. Krupski, & E. Shriberg, "Labeling Guide for Dialog Acts in the ICSI Meeting Recorder Meetings", ICSI Technical Report, 2002.

[2] G. Bruce et al., "On the analysis of prosody in interaction", in *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pp. 43–59, Springer: New York, 1997.

[3] J. Carletta, "Assessing agreement on classification tasks: The Kappa statistic", *Computational Linguistics*, 22(2), pp. 249–254, 1996.

[4] J. Carletta et al., "The coding of dialogue structure in a corpus", in J. Andernach et al. (Eds.) *Proc. Ninth Twente Workshop on Language Technology: Corpus-based Approaches to Dialogue Modelling*, Univ. Twente, Enschede, 1995.

[5] M. Finke et al., "Clarity: Inferring discourse structure from speech", in J. Chu-Carroll & N. Green (Eds.), *Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium*, pp. 25–32, AAAI Press, Menlo Park, 1998.

[6] J. Hirschberg & C. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues", *Proc. ACL*, Santa Cruz, pp. 286–293, 1996.

[7] A. Janin et al., "The ICSI Meeting Corpus", *Proc. ICASSP*, Hong Kong, 2003 to appear.

[8] D. Jurafsky, E. Shriberg, & D. Biasca, "Switchboard-DAMSL Labeling Project Coder's Manual", Tech. Report 97-02, U. Colorado Institute of Cognitive Science, 1997.

[9] M. Mast et al., "Dialog act classification with the help of prosody", *Proc. ICSLP*, vol. 3, pp. 1732-1735, Philadelphia, 1996.

[10] E. Shriberg et al., "Can prosody aid the automatic classification of dialog acts in conversational speech?", *Language and Speech*, 41(3-4), pp. 439-487, 1998.

[11] E. Shriberg, A. Stolcke, D. Hakkani-Tur, & G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics", *Speech Communication*, 32(1-2), pp. 127-154, 2000.