THE RELATIONSHIP BETWEEN DIALOGUE ACTS AND HOT SPOTS IN MEETINGS

Britta Wrede[1,2] *Elizabeth Shriberg*[1,3*]

[1]International Computer Science Institute, Berkeley, USA
[2]Applied Computer Science Group, Bielefeld University, Germany
[3]Speech Technology and Research Laboratory, Menlo Park, USA
bwrede@techfak.uni-bielefeld.de ees@speech.sri.com

ABSTRACT

We examine the relationship between hot spots (annotated in terms of involvement) and dialogue acts (DAs, annotated in an independent effort) in roughly 32 hours of speech data from naturally-occurring meetings. Results reveal that four independently-motivated involvement categories (non-involved, disagreeing, amused, and other) show statistically significant associations with particular DAs. Further examination shows that involvement is associated with contextual features (such as the speaker or type of meeting), as well as with lexical features (such as utterance length and perplexity). Finally, we found (surprisingly) that perplexities are similar for involved and Non-involved utterances. This suggests that it may not be the amount of propositional content, but rather participants' attitudes toward that content, that differentiates hot spots from other regions in a meeting. Overall, these specific correlations, and their relationships to other features such as perplexity, could provide useful information for the automatic archiving and browsing of natural meetings.

1. INTRODUCTION

Whether we like it or not, meetings have become a ubiquitous part of everyday life for many professionals. Keeping track of what went on in a meeting can be a difficult task, especially since taking notes competes with the cognitive resources necessary for active participation in the meeting. Although the audio (and video) record of a meeting can be used for archival purposes, reviewing the raw recording in real time is typically too tedious to be practical. What is necessary is some means to automatically browse and retrieve locations of interest–for example, regions in which participants disagreed, came to a decision, changed topics, and so on. To this end, recent efforts in speech technology have focused on automatically transcribing the words spoken in a meeting, as well as on capturing discourse information and other potentially useful patterns in speakers' behavior [1, 2, 3, 4].

One important capability for such a technology would be a way to detect "hot spots". By hot spots, we mean regions in which participants are highly involved in the discussion—regardless of the nature of that involvement. We will use the term *involvement* to refer to what is also described in the literature on emotion as *activation*, or the "strength of a person's disposition to take action" [5].

Prior work found that subjects could reliably rate involvement at the utterance level in natural meetings. Furthermore, ratings of higher involvement were correlated with higher values for automatically extracted prosodic features, including pitch and intensity [6]—consistent with patterns observed for higher levels of activation in the emotion literature [5]. While such work suggests that involved utterances can be identified and are distinguished prosodically, little is known about the discourse-level, pragmatic or semantic nature of hot spots. We explore this question in the present paper, by looking at the relationship between annotations of perceived involvement and independent annotations of dialogue acts.

Dialogue acts (DAs) reflect the functions that utterances serve in a discourse. For example, utterances can serve as statements, questions, or acknowledgments of another speaker's contributions. They often incorporate semantic and/or syntactic information. DAs have been extensively analysed in different contexts, such as scheduling tasks [7] or telephone speech [8]. The meeting context, however, is a relatively new domain for DA annotation [9].

2. METHOD

2.1. The Meeting Recorder Corpus

The analysis is based on the ICSI Meeting Recorder corpus [10, 2], which consists of naturally occurring meetings of different research groups at ICSI. The meetings cover sci-

*This work was supported by the German Academic Exchange Service (DAAD) through a Postdoctoral Fellowship, and by a DARPA Communicator project, ICSI NSF ITR Award IIS-0121396, SRI NSF IRI-9619921, SRI NASA Award NCC2-1256, and the Swiss National Science Foundation through the research network IM2.

1. General Tags:		3. Specific Tags:			
s	statement	2	collabor.	df	defend./expl.
sj	subjective s	aa	accept	e	elaboration
qh	rhethor. q.	aap	partial acc.	f	follow me?
qo	open quest.	am	maybe	fa	apology
qr	Or quest.	ar	reject	fe	exclam.
qrr	Or-cl.-q	arp	partial rej.	ft	thanks
qw	Wh-quest.	ba	assessment	g	tag quest.
qy	Y/N quest.	bc	correct oth.	j	joke
b	backchan.	bd	downplayer	m	mimic
fg	floor grab.	bk	acknowl.	na	affirm. ans.
fh	floor hold.	bs	summary	nd	dispref. ans.
h	hold	bsc	self-corr.	ng	negative ans.
2. Disruptions:		bu	understand?	no	uncert. resp.
%	indeciph.	cc	commitm.	r	repeat self
%-	interrupt.	co	command	t	about task
%–	abandoned	cs	suggest.	t3	3rd party
x	nonspeech	d	declar. q.	tc	topic chg.

Table 1. DA labels that were used for the anotation of the Meeting Recorder data and their short descriptions. For a more detailed description cf. [9], adapted from [11].

entifi c topics as well as more administrative topics like machine usage or the scheduling of other meetings. The current analysis examined 32 meetings, each about one hour long. The number of participants per meeting ranged from 5 to 8.

2.2. Annotation of DAs

DA annotations were based on a previous approach [11], but adapted for meetings in a number of ways as described in [9]. Table 1 shows the DA labels used. A DA consists of exactly one general tag, which describes the type of utterance (e.g. statement, question or backchannel), plus a variable number of tags that describe the pragmatic function of the segment such as the turn-taking function (e.g. floor holder, floor grabber) or the illocutionary act (e.g. accept, reject, commit). For example:

SPK	Words	DA label
A:	they're doing some noise removal thing.	< s >
A:	right?	<qy ^ d ^ g>
B:	yeah yeah.	< s ^ aa>

2.3. Annotation of Involvement

Importantly, the annotation of involvement was independent of the annotation of DAs. The two different types of annotation were performed by different labelers, for different research projects. The annotations of involvement were performed by one rater while listening to whole meetings. The involvement class was further divided into three subclasses: amused, disagreeing and other. The fi rst two subclasses were identifi able and different from each other; the third class was used for all remaining cases.

	Involved	Non-involved	Total
M-segs	811 (2.2%)	38,468 (97.8%)	39,279

Table 2. Frequencies of involved and non-involved M-segs over all 32 meetings.

	Involved		Non-involved	
Case	I-Segs	M-Segs	I-Segs	M-Segs
1	223 (47.3%)	223 (27.5%)	800	800
2	168 (35.7%)	448 (55.2%)	-	-
3	70 (14.9%)	124 (15.3%)	-	-
4	10 (2.1%)	16 (2.0%)	-	-
Sum	471 (100%)	811 (100%)	800	800

Table 3. Frequencies of all involved and non-involved I-segs and M-segs used in this study according to alignment cases. Refer to text for a description of the cases.

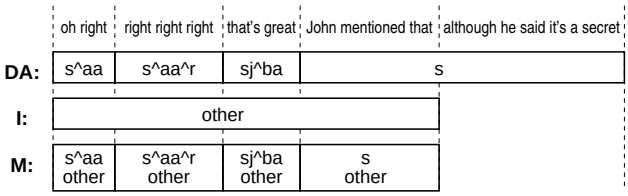

Fig. 1. Segmentation and alignment of an utterance with different boundaries for DA-segs and I-segs and the resulting M-segs.

Utterances were labeled as involved if the rater perceived a higher level of affect in a speaker's voice than was typical for that speaker in the meeting. For example, the speaker could sound especially interested, surprised or enthusiastic about what is being said, or he or she could express strong disagreement, amusement, or stress. We found in earlier work that although this task is subjective and raters do not always feel confi dent in their judgements, listeners actually agree surprisingly well [6].

2.4. Segmentation and Alignment

As noted earlier, annotation of involvement was done separately from annotation of DAs. In each effort, a method of segmentation was developed, appropriate to the task at hand. Thus, DA segments and involvement segments did not always coincide. In order to align DA segments (DA-segs) and involvement segments (I-segs), four cases were defi ned; these accounted for almost all of the actually occurring cases.

In the fi rst case, the boundaries of an I-seg match exactly those of the corresponding DA-seg. In the second case, one I-seg comprises several DA-segs. Here, the DA-tags of all internal DA-segs are taken into account for the analysis. In cases 3 and 4, only the fi rst (3) or the last (4) boundary matches the corresponding I-seg boundary. In these cases

DA-tags (# M-segs)	non-inv (800)	other (499)	amused (174)	disag (138)
s	− 63.2	69.9	73.0	76.1
b	++ 17.5	−− 1.8	−− 2.3	−− 0.7
fh	+ 4.8	− 1.4	2.9	2.2
bk	+ 5.2	2.8	1.1	2.9
sj	−− 2.1	++ 10.4	7.5	3.6
df	−− 1.2	++ 6.0	1.7	4.3
fg	− 1.8	++ 5.6	1.7	2.9
ba	1.7	++ 5.0	4.0	−− 0.0
bsc	− 0.0	++ 1.4	0.0	0.7
r	0.5	+ 2.0	0.6	0.0
fe	0.5	+ 1.8	0.6	0.7
j	−− 0.1	−− 0.2	++ 21.8	0.0
cs	−− 2.7	+ 6.6	++ 9.8	2.2
d	3.1	3.6	+ 6.9	3.6
ar	−− 0.4	2.2	0.6	++ 10.1
nd	0.9	1.2	1.1	++ 8.7
ng	0.2	0.2	0.0	++ 5.1
arp	0.2	1.2	0.0	++ 2.9
aap	0.2	0.2	0.0	++ 2.9

Table 4. Percentages of M-segs in which a certain DA-tag occurred. Statistical signficance (Chi-square test) is denoted by $++/--$ for $p < .01$ and $+/-$ for $p < .05$. The $+/-$ signs indicate significantly higher or lower percentages, respectively, than the null hypothesis.

the DA-seg has to be split, which could be problematic since it would involve altering DA tags. As it turned out, however, splits occurred only within DAs that were labeled as *s* ('statement'), and inspection revealed that the split occurred at a clause boundary. Thus, each part of the split could constitute a statement on its own. Such an example occurs in Fig. 1: *"John mentioned that"* and *"although he said it's a secret"* are each rewritten as an individual statement. Fig. 1 also shows an example where the end of an I-seg falls within a DA-seg. To handle such a split, a new type of segment was introduced: the *M-seg*. This unit restarts when encountering either a DA-seg boundary or an I-seg boundary; it thus represents the maximal overlap of DA-segs and I-segs.

Table 2 shows the number of segments that were labeled and included in the data. Rates of the various cases are shown in Table 3. Even though M-segs allow for DA splits, most of the observed cases coincide with DA-segs. This suggests that involvement and DAs occur on similar time scales.

2.5. Additional Features

Contextual features. Although our focus was on the relationship between involvement and DAs, we also examined two contextual features: speaker identity and meeting type. Analyses showed that both were correlated with involvement—particularly speaker identity. However, because our data included only a small set of speakers who spoke often, and even fewer who produced involved utterances, it was not clear that analyses on this factor would generalize. Thus, only descriptive statistics for these features are presented here.

Lexical features. Another obvious potential correlate of involvement is lexical information. We examined two simple features: the length of the M-seg in words, and perplexity of the M-seg. Perplexity is a measure of information content; it can be interpreted as the average number of possible words following any word. Perplexity was estimated using a 4-gram language model trained on Switchboard, English CallHome, Broadcast News, and web data.

3. RESULTS

3.1. Association between DAs and Involvement

As a first step in understanding the relationship between involvement and DAs, the relative frequency of each DA tag was computed over each of the four involvement classes (non-involved, amused, disagreeing, other). To compute relative frequencies, absolute counts were normalised by the number of M-segs. Since each DA can consist of more than one tag, the sum of all frequencies can be higher than 1.0. Results are shown in Table 4.

A number of interesting observations deserve mention. First, other involvement exhibits significantly more subjective statements (sj), defending or explaining statements (df), floor grabbers (fg), assessments or appreciations (ba), and self-corrections (bsc). It has slightly more suggestions (cs), repetitions (r) and exclamations (fe) than the other groups pooled together. This suggests that other involvement is characterised by more subjective and evaluative contributions (sj, df, ba, cs). The high rate of floor grabbers may also indicate an unusual pattern of initiative shifts at the discourse or task level ([12]), and merits further investigation.

Second, amusement is often, but not always, characterised by jokes (j). It is also associated with suggestions (cs) and declarative questions (d). Inspection revealed a close relationship between jokes and suggestions in amused utterances, indicating that information that is not meant seriously is not always marked as such. On the other hand, amusement often occurs in the absence of jokes and may not necessarily mean that information is not meant seriously.

Third, disagreement is characterised by significantly more negative responses in the form of rejections (ar), negative (ng) and dispreferred answers (nd), as well as partial rejections (arp) and partial acceptances (aap). In our meeting data, these DAs are very rare and are also viewed as quite impolite. We found that differences of opinion in our data are generally conveyed by positive statements rather than open contradictions. The disagreement class thus comprises the very strongest and tense disagreements.

Finally, in contrast to the former three cases, non-involved utterances can be best described as lacking in

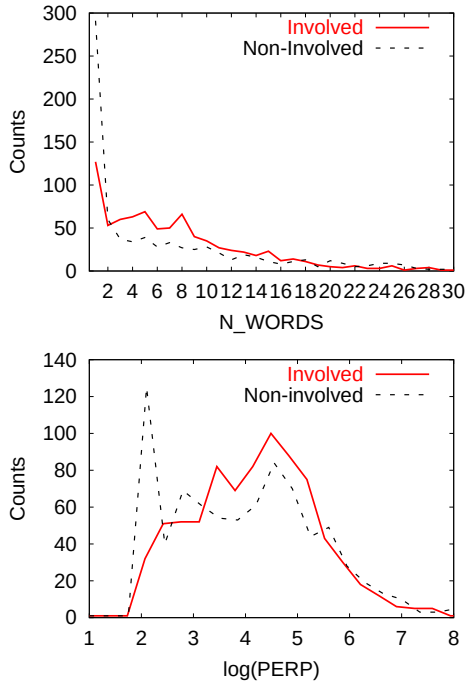

Fig. 2. Histograms of Number of words (N-WORD) and the Perplexity (PERP) of the analyzed involved and non-involved M-segs.

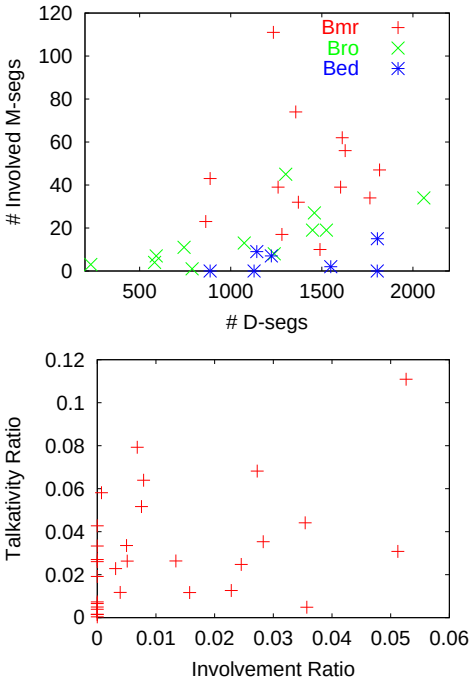

Fig. 3. Meetings plotted by their number of involved and non-involved utterances (top). Speakers plotted by their talkativity and involvement ratio.

specifi c DA characteristics (sj, cs, df, j, fg, ar). Non-involved utterances are, as one would expect, correlated with backchannels (b, bk). Quite interestingly, however, non-involved utterances have signifi cantly higher rates of floor holders (fh) than involved utterances, even though one might assume the opposite. This may be because in less involved regions, the floor is less at stake and speakers have time to pause without losing the floor. They thus insert more floor holders because they make more pauses. This hypothesis could be investigated by examining associated pause distributions and turn-taking behavior in both involved and non-involved speech.

The distributions for the N_WORDS (number of words) features (Fig. 2) reveal a marked difference between involved and non-involved M-segs. The peak for one-word utterances in the non-involved curve results mainly from the high number of backchannels in this class. Another noteworthy fi nding is that the distributions for perplexity show almost no difference between involved and non-involved speech (other than the expected peak due to backchannels for non-involved speech). This is somewhat surprising. It suggests that hot spots may not contain more *unusual propositional* information than the rest of the conversation, but rather more *pragmatically "loaded"* information.

We also looked at the distribution of involvement with respect to contextual factors, including the meeting type and speaker differences. The top of Figure 3 shows the numbers of involved against all utterances in each meeting. As shown, overall rates of involvement are associated with meeting type, since for example 'Bmr' meetings tend to have more involvement than 'Bro' meetings. Bro meetings were guided by a moderator asking for reports while the Bmr meetings were more casual and less hierarchically structured. Such factors are likely to influence the involvement patterns of the participants.

We also wondered how involvement varies by speaker. In addition to overall rates, we were also interested in whether speakers who tend to talk more also tend to be more involved. We looked at the relative involvement of a speaker (computed as the number of involved M-segs of a speaker normalized by his or her total number of D-segs) against his or her talkativity ([4], here computed as the ratio of the total number of D-segs of that speaker divided by the total number of D-segs that occurred in all meetings that the speaker attended). Results, as shown in Fig. 3, suggest that overall talkativity and involvement are not correlated. This is interesting; it suggests that individual speakers may have different "involvement" thresholds for deciding when to talk. (Perhaps meetings would be more effi cient if people raised their thresholds for deciding when to speak).

3.2. Prediction of Involvement by DAs

In addition to examining correlations of individual DAs with involvement, we asked which DAs are most useful in jointly predicting involvement. For example, jokes (j) are associated with amusement but they are not suffi cient to predict amusement since they only occur in about 22% of the

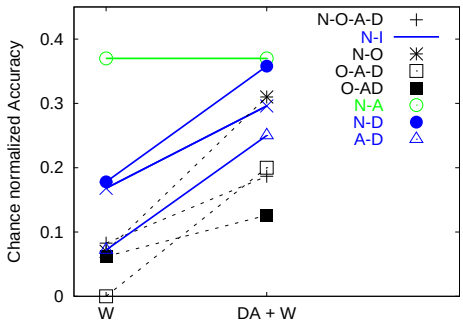

Fig. 4. Chance-normalised performance of classification based on word information only (W) versus DA and word information (DA+W), for the eight classification tasks described in Section 3.2.

amused M-segs.

We used CART-style decision trees as classifiers to predict involvement in unseen test data [13]. Because of the high skew in sample size of classes in some comparisons, we resampled data to equate class priors. Resampling was limited to a factor of 3. The data was randomly divided into a test set of 322 M-segs and a training set of 1289 M-segs. In order to analyse which features distinguish which classes, eight different meaningful tasks were defined. The tasks were: (1) N-O-A-D, (2) N-I, (3) N-O, (4) O-A-D, (5) O-AD, (6) N-A, (7) N-D, and (8) A-D, where '-' denotes a contrast, 'N' denotes non-involved, 'O' other, 'A' amused, 'D' disagreeing and 'I' involved (= pooled over all classes OAD). 'AD' denotes the pooled data set of 'A' and 'D'.

Additionally, we used two different feature sets: one with only lexical information (W), and one with both lexical and DA information (DA+W). Fig. 4 shows a summary of the results. In order to compare across conditions differing in class size (and hence, in chance performance) a normalised value is computed as $(p_C - p_E)/(1 - p_E)$ where p_C denotes the percentage of correctly classified items and p_E the expected percentage. Note that actual accuracies are much higher than these chance-normalized values.

In the graph, each line represents results for one of the eight classification tasks. As can be seen, adding DAs significantly increases the classification accuracy for all but one task. The special case is the task of discriminating amused from non-involved M-segs (N-A). It turned out that in this case, no DA features were used for the classification, because perplexity and word count alone provided best results. Although this result merits further investigation, an inspection of the data suggests that some amused utterances have unusually high perplexities because of the use of unusual words or word sequences. For example, the following example occurred in a discussion about whether people would lie when asked about their age in filling out consent forms for the meeting recording project:

"Jack Benny was thirty-nine for forty years"

It is worth mentioning that this perplexity effect for amused speech could become even more pronounced if we estimated perplexities based on in-domain ICSI meeting data (once enough data were available) rather than on the broader range of domains we used here.

To better understand the differences between involvement classes, we looked at the features that contribute most to specific class discriminations. Fig. 5 indicates the contribution that different DAs make for the eight classification tasks. In all cases the results are from the condition in which only DA features were available to the classifer. The measure on the ordinate, "Feature Use", reflects the proportion of samples in the decision tree that are affected by querying a particular feature. Feature use thus sums to 1.0 for each experiment.

The first graph shows that for the distinction between involved and non-involved M-segs (N-O-A-D and N-I), as well as for the distinction among the involved classes, features that are generally relevant for classification contribute roughly equally.

When distinguishing between non-involved M-segs and each of the involvement classes alone, different features become important. In the case of N-D, we see that not only the negative response tags (nd and ar) and the characteristic features for non-involvement (b, bk, fh) are used, but also the acceptance feature (aa) which did not show up as significant in the descriptive analysis of non-involvement. A particularly interesting question is what distinguishes the involvement classes from each other. Fig. 4 shows that the O-A-D and O-AD tasks use quite different features. For distinguishing A-D, the features that are relevant for disagreement (nd, ar) and jokes (j) are used. However, features that are more relevant to involvement in general are also selected (fh, sj, cs). This suggests that in the absence of other characteristic features for amusement, the more general characteristics of involvement become important.

Given these significant correlations between involvement classes and DAs it would be interesting to also incorporate sequence information into the classification process. In [14] it has been shown that different types of discourse can be distinguished by modelling sequences of conversational acts with HMMs. One might expect that the sequencing provides additional information. However, due to the high frequency of overlap in multi-party conversations sequence information is difficult to determine.

4. CONCLUSION

We found that annotations for four involvement categories (non-involved, disagreeing, amused, and other) show statistically significant associations with independently-annotated DAs. While some results were expected (such as the correlation between jokes and amusement, or backchannels and non-involvement), others were surprising (such as

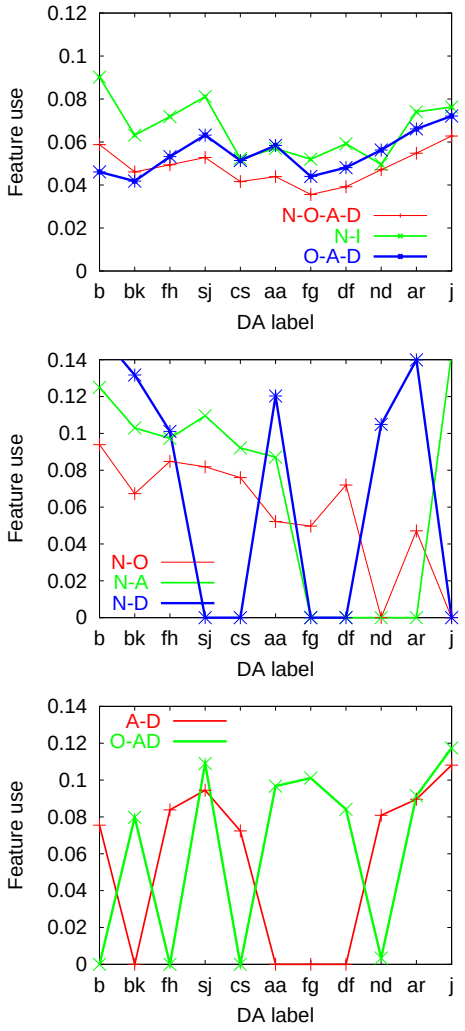

Fig. 5. Relative contribution of different DA tags used for the discrimination of involvement classes. Feature use is normalized (sums to 1) and is based on number of tokens affected by querying the feature.

the correlation between floor holders and non-involvement). Further examination showed that involvement is associated with contextual features (speaker, type of meeting), as well as with lexical features (utterance length, perplexity). For many specifi c involvement contrasts, a combination of lexical and DA features yielded better prediction than lexical information alone. An interesting exception, however, was the contrast between amusement and non-involved utterances; in this case lexical features alone were the best predictors, suggesting that amused utterances tend to contain unusual words or word sequences. Finally, with the exception of a difference caused by frequent one-word backchannels in the non-involved class, we found (rather surprisingly) that perplexities are similar for involved and non-involved utterances. This suggests that it may not be the propositional content that differs in the two cases, but rather the affective response of meeting participants to that content. Overall, these specifi c correlations, and their relationships to other features such as perplexity, could provide useful information for the automatic archiving and browsing of natural meetings.

5. REFERENCES

[1] "Rich transcription 2002 evaluation," http://www.nist.gov/speech/tests/rt/rt2002/.

[2] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "Meetings about meetings: research at ICSI on speech in multiparty conversations," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, Apr. 2003.

[3] A. Waibel, H. Yu, M. Westphal, H. Soltau, T. Schultz, T. Schaaf, Y. Pan, F. Metze, and M. Bett, "Advances in meeting recognition," in *Proc. Int. Conf. Human Language Technology Research*, James Allan, Ed., San Diego, 2001, pp. 11–13.

[4] S. Renals and D. Ellis, "Audio information access from meeting rooms," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[5] R. Cowie, E. Couglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[6] B. Wrede and E. Shriberg, "Spotting 'hot spots' in meetings: Human judgments and prosodic cues," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, 2003, pp. 2805–2808.

[7] E. Noeth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann, "On the use of prosody in automatic dialogue understanding," in *Proc. DIAPRO*, 1999, pp. 25–34.

[8] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 3, no. 26, pp. 339–373, 2000.

[9] S. Bhagat, R. Dhillon, H. Carvey, and E. Shriberg, "Labeling guide for dialog act tags in the meeting recorder meetings," Tech. Rep. 2, International Computer Science Institute, Berkeley, August 2003.

[10] A. Janin, D. Baron, J. Edwards, E. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003.

[11] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard-DAMSL Labeling Project Coder's Manual," Tech. Rep. 97-02, University of Colorado, Institute of Cognitive Science, Boulder, CO, 1997, `http://www.colorado.edu/ling/jurafsky/-manual.august1.html`.

[12] J. Chu-Carroll and S. Carberry, "Conflict resolution in collaborative planning dialogs," *Int. J. Human-Computer Studies*, vol. 53, pp. 969–1015, 2000.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Pacifi c Grove, CA, 1984.

[14] A. Soller and P. Busetta, "An intelligent agent architecture for facilitating knowledge sharing communication," in *Proc. Workshop on Humans and Multi-Agent Systems at Int. Conf. on Autonomous Agents and Multi-Agent Systems*, Melbourne, 2003, pp. 94–100.