COMPARING THE CONTRIBUTIONS OF CONTEXT AND PROSODY IN TEXT-INDEPENDENT DIALOG ACT RECOGNITION

Kornel Laskowski [1] *Elizabeth Shriberg* [2,3]

[1] Carnegie Mellon University, Pittsburgh PA, USA
[2] SRI International, Menlo Park CA, USA
[3] International Computer Science Institute, Berkeley CA, USA
kornel@cs.cmu.edu ees@speech.sri.com

ABSTRACT

Automatic segmentation and classification of dialog acts (DAs; e.g., statements versus questions) is important for spoken language understanding (SLU). While most systems have relied on word and word boundary information, interest in privacy-sensitive applications and non-ASR-based processing requires an approach that is text-independent. We propose a framework for employing both speech/non-speech-based ("contextual") features and prosodic features, and apply it to DA segmentation and classification in multiparty meetings. We find that: (1) contextual features are better for recognizing turn edge DA types and DA boundary types, while prosodic features are better for finding floor mechanisms and backchannels; (2) the two knowledge sources are complementary for most of the DA types studied; and (3) the performance of the resulting system approaches that achieved using oracle lexical information for several DA types. These results suggest that there is significant promise in text-independent features for DA recognition, and possibly for other SLU tasks, particularly when words are not available.

Index Terms— Dialog act tagging, Prosody, Turn taking, Speech activity modeling, Privacy-sensitive features, Meetings.

1. INTRODUCTION

The past decade has seen rapid improvement in systems that automatically segment a spoken conversation into dialog acts (DAs; e.g., statements versus questions) [1, 2, 3, 4]. Such systems can inform downstream language processing applications, including dialog systems, summarization, and machine translation [5]. A key knowledge source in DA recognition (or joint segmentation and classification) systems is lexical information: word boundaries define the possible locations of DA boundaries, and specific word sequences correlate with particular DAs (for example, "wh-" words in English correlate with questions, and "uh-huh" is typically a backchannel). Human annotators appear to rely heavily on word identity when segmenting and labeling DAs.

An interesting question thus arises when one needs DA recognition, but words are not available. This is the case in a growing line of work exploring "privacy-sensitive" data (such as [6, 7]), in which the actual content of a conversation cannot be revealed. Words are also unavailable in situations in which automatic speech recognition (ASR) is too expensive (e.g., given the amount of data, or need for speed). Although there is little work on the task of text-independent DA recognition to date, a recent study found that simple speech/non-speech features for multiple participants provided significant information for DA recognition in natural meetings [8]. We refer to such features as "contextual", since they capture the relationship of stretches of talk ("talkspurts") across speakers.

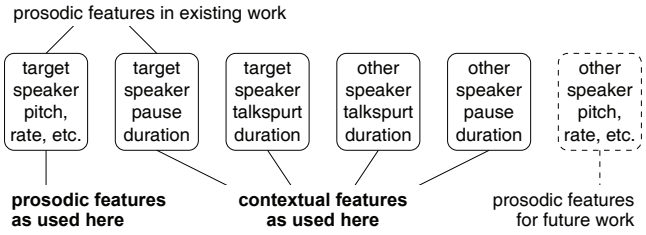


Fig. 1. "Word-free" feature types and their grouping in traditional work on prosody for DA recognition, as well as the grouping employed in the current work.

In this work we consider a second source of text-independent information, namely prosodic information. While there is a long history of work on prosody and discourse, we know of no previous work that explores completely word-independent prosodic information for the task of automatic DA recognition. Our investigation seeks answers to three basic questions:

1. *How much does context versus prosody contribute to text-independent DA recognition?*
2. *To what extent are context and prosody complementary and does this depend on the DA type?*
3. *How do these text-independent systems compare to a system that uses the words?*

In comparing feature types, we note that one unusual aspect of our feature classes is that pause information, which is typically grouped with prosodic features, is already captured by speech/non-speech information. Because of this, our prosodic feature set includes only beyond-pause features, such as pitch, energy, and speaking-rate-related information (and any improvements from prosody are thus interpreted as coming from features other than pause). A second aspect of our work is that we model speech/non-speech information from not only the target speaker but also from other talkers (we have not yet extended prosodic information in this way). For a clarification of differences between our feature set classes and those used in many other systems, see Figure 1. A third characteristic of our approach is that we use a novel but simple hidden Markov model (HMM) framework as a means by which to evaluate and compare the contributions of context and prosody.

The framework is intentionally extensible to modeling any feature type, across multiple speakers, and can be easily re-trained for other domains or corpora. An HMM framework was chosen to facilitate future integration within a realtime acoustic decoder.

2. DATA

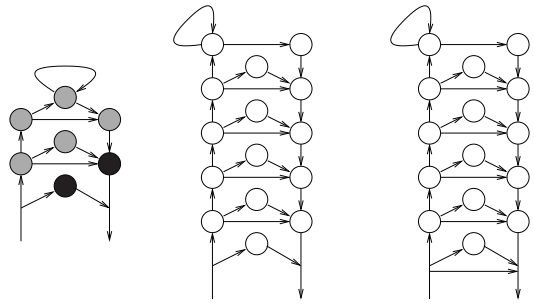
We study data from the ICSI Meeting Corpus, consisting of 75 longitudinal recordings of naturally occurring meetings by several groups at ICSI [9, 10]. We use the previously published split of this data [10] into a TRAINSET of 51 meetings, and a DEVSET and a TESTSET of 11 meetings each, without additional splits, to facilitate comparison with the work of others. The meetings are provided with lexical forced alignment and DA annotation. We focus on three groups of DA types. The first is that of floor mechanisms, and includes floor grabbers (`fg`), floor holders (`fh`), and holds (`h`). The second group consists of backchannels (`b`) and acknowledgments (`bk`); we also consider accepts (`aa`). All six have been reported to share a common vocabulary [10], suggesting that lexical content may not adequately distinguish among them. All other speech implements either statements (`s`) or questions (`q`), representing propositional-content DAs; these terminate in one of three ways: they can be completed (*com*), interrupted (*int*), or abandoned (*aba*). The distribution over DA type (see Table 1 of Section 4) is significantly skewed towards statements, as is typical for most conversational corpora.

3. SYSTEMS

Our Viterbi decoder operates at a frame size and step of 100 ms; a larger frame step would prevent us from correctly identifying very short DAs. The state topology models the sequencing of DA production in time, including the duration of talkspurts implementing the 8 DA types of interest, the duration of non-speech gaps between DAs, and the duration of non-speech gaps internal to DAs. These durations are modeled *implicitly*; explicitly, we model only the sequence of frame-level speech/non-speech posteriors, as might be provided by a speech activity detector. The sub-topology for each DA type consists of a punctuation-bearing talkspurt fragment, an optional non-punctuation-bearing talkspurt fragment, and an optional intra-DA non-speech gap; the former two elements have near-identical structure and are shown in Figure 2(a), while the latter is shown in Figure 2(b). Each DA sub-topology is connected to every other, independently in both directions, via an inter-DA gap sub-topology shown in Figure 2(c). The resulting complete HMM topology[1] implements a general network which allows for the splitting and merging (across intra-DA gaps) of talkspurts into DAs. All transition probabilities are estimated from the best forced alignment path in TRAINSET.

3.1. Contextual Features & Baseline System

We model the speech/non-speech activity for other (non-"target") participants in feature space. In [8], we found that a 10-second-window snapshot of the speech activity of the *locally most talkative* other participant(s) improves retrieval F-scores for most of the DAs studied, sometimes quite dramatically. The computation of this snapshot consists of: (i) sampling the speech/non-speech posterior of the target speaker and of her/his locally most talkative interlocutor(s), by tiling the 10-second context window with 0.5-second tiles; and (ii) retaining the top 24 discriminants following rotation via linear discriminant analysis (LDA). Because we are interested in inherent characteristics rather than a particular speech activity detection system, we use reference as opposed to automatically-derived segmentation. The rotated features are modeled with a 64-element state-specific Gaussian mixture model (GMM), and the log-likelihood is combined linearly with the state transition log-probability using a weight λ_{Cont}. The feature transformation and model complexity were tuned to optimize the unweighted average F-score over the 8 DA types using DEVSET.



(a) talkspurt fragment (b) intra-DA gap (c) inter-DA gap

Fig. 2. Subnetworks in an HMM topology for conversational speech, with a frame step of 100 ms; states shown in white denote non-speech, while all other states denote speech. In (a), egress states, which bear punctuation in punctuation-bearing talkspurt fragments, are shown in black. Note that (b) and (c) are identical, except that inter-DA gaps may have zero duration.

3.2. Prosodic Features & System

As mentioned earlier, pause information is already captured by our contextual system; here, we explore other correlates of prosody. For intonation, we estimate the 7-element-filterbank fundamental frequency variation (FFV) representation [11]. FFV features capture the per-frame change in F0, are simpler to compute and model than standard pitch estimates, and do not require estimation of base pitch for speaker normalization. Loudness, voice quality, and speaking rate are captured using per-frame log-energy, the first-order difference of log-energy, the normalized first off-axis autocorrelation maximum, and the first-order cosine difference of Mel-filterbank responses and of Mel-filterbank log-responses. These 12 features are computed at a frame step of 8 ms and a frame size of 32 ms, as in [11]; 12.5 consecutive feature vectors on average are aligned to each 100 ms decoder frame. Eleven discriminants are retained following rotation via LDA. Emission probabilities are modeled with a state-specific 32-element GMM, and a model weight λ_{Pros}, corresponding to λ_{Cont} in Subsection 3.1, is used during decoding.

To compare the contributions of context and prosody, we also construct the model-space combination of the two systems, reoptimizing only λ_{Cont} and λ_{Pros}, using DEVSET.

3.3. Lexical Bigram Features & System

We compare the results of our text-independent systems to a state-of-the-art oracle lexical system. We note that this system is hard to beat because (1) it relies on words, which were used by annotators

[1] DA sub-topologies for propositional-content DAs consist of $7 + 15 + 3\times7 = 43$ states; all others DA sub-topologies consist of only $7+15+7 = 29$ states. The total number of states is $2\times43+6\times29+8\times8\times15 = 1220$.

		Topo only	Context (Cont)		Prosody (Pros)		Cont & Pros		Bigrams (Lex)	
		(g=c)-Opt	g-Opt	c-Opt	g-Opt	c-Opt	g-Opt	c-Opt	g-Opt	c-Opt
DA Types										
mean	prior	21.8	29.3	31.1	31.5	33.7	38.4	39.8	53.0	54.5
floor holder, `fh`	2.7%	11.3	24.0	25.6	37.7	39.5	43.5	43.7	62.3	63.5
hold, `h`	0.3%	†0.0	8.5	†6.3	25.0	17.1	31.8	‡29.2	33.9	‡41.5
floor grabber, `fg`	0.6%	0.0	12.5	†13.7	7.2	7.2	11.6	†14.0	24.5	24.5
backchannel, `b`	2.8%	†57.1	54.7	†57.8	48.0	64.6	64.5	66.9	77.0	77.0
acknowledgment, `bk`	1.5%	3.2	15.7	14.9	19.0	20.9	24.2	25.6	56.3	56.3
assert, `aa`	1.1%	2.6	12.3	†13.0	9.5	8.9	14.0	†16.0	38.1	40.0
statement, `s`	84.5%	†91.4	82.3	†91.3	85.8	‡91.8	87.3	‡91.8	91.9	93.3
question, `q`	6.6%	8.8	23.9	26.3	19.6	19.6	30.4	30.9	39.8	39.8
DA Termination Types										
completed, *com*		53.1	58.3	62.1	59.1	59.1	63.4	63.7	68.0	69.1
interrupted, *int*		0.0	22.6	22.6	10.5	11.8	26.0	28.7	21.9	21.9
abandoned, *aba*		0.0	6.6	†6.6	2.4	3.6	5.4	†7.6	11.4	13.0

Table 1. EVALSET F-scores for all 8 DA types, including their arithmetic mean, and for 3 DA termination types. Performance is shown for five systems: the topology alone; context features in the topology; prosodic features in the topology; model-space combination of the context and the prosodic systems; and the lexical system. "g-Opt" and "c-Opt" represent globally optimized and condition-optimized variants, respectively. "†" and "‡" identify within-row "c-Opt" pair members whose difference is not statistically significant; differences for all other within-row "c-Opt" pairs are significant at the $p < 0.005$ level.

to segment and label DAs, and (2) because it uses true rather than automatically recognized words.

The oracle system used here employs the same HMM topology employed by our other systems. The observables in each speech state are the left and right bigram of the word whose timespan coincides with the 100 ms duration of the state in question. Where words are separated by more than 0.7 s of non-speech, we insert a `SIL` token; bigrams are defined over a vocabulary consisting of: the `SIL` token, those words whose frequency of occurrence exceeds 0.1% for any of the 8 DA types of interest, and the `UNK` token to which all other words are mapped. The log-likelihoods of the left and right context are equally weighted, and combined with the state transition log-probability using a weight λ_{Lex}. This system was found to perform slightly better [8] than the hidden event language model on a more standard 5-DA-type classification task [1].

4. RESULTS AND DISCUSSION

We compare the performance of our 5 systems, including a topology-only system ("Topo only") in which the feature space consists of only the speech/non-speech posterior of the target speaker at time t. The remaining 4 systems are the contextual system ("Cont"), the prosodic system ("Pros"), the combined contextual and prosodic system ("Cont & Pros"), and the lexical bigram system ("Lex"). Performance for all 5 is measured using the F-score over 100 ms speech frames; we do not measure F-score over DA *units*, since DA segmentation is not given and must be inferred simultaneously.

Table 1 shows results for EVALSET for two conditions: g-Opt, in which the emission model weight λ is tuned using DEVSET to maximize the *average* 8-DA F-score, and c-Opt, in which that same weight is optimized using DEVSET for the *particular* DA type or DA termination type. As shown, g-Opt yields mean F-scores over the 8 DA types which are 1.4-2.2% lower on EVALSET than when they are computed over the 8 c-Opt variants. This is on par with the same difference for DEVSET (not shown). However, for some DA types (e.g. holds), the c-Opt F-score is lower than in the corresponding g-Opt score, which is a case of overfitting to DEVSET. This is observed only for DA types which are infrequent by time (as shown next to each DA in the first column).

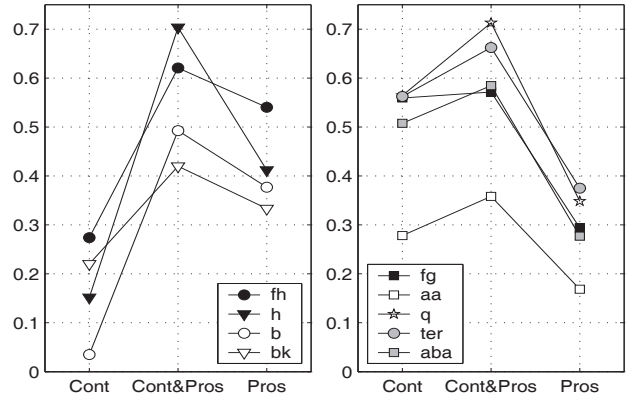
For all 11 c-Opt conditions in Table 1, we also assess the statistical significance of the observed difference in F-score among the five systems[2]. Six differences were analyzed: between "Topo only" and both "Cont" and "Pros", between "Cont" and "Pros", between both "Cont" and "Pros" and "Cont & Pros", and between "Cont & Pros" and "Lex". With only 8 exceptions, all 66 pairwise comparisons were found to be significant at the $p < 0.005$ level (we note the exceptions, shown using "†" and "‡" in Table 1, in what follows).

Several observations allow us to answer the questions in Section 1. First, unsurprisingly, contextual features are the best single feature type for detecting whether DAs are interrupted (*int*), outperforming even lexical features. Although prosodic features are relatively poor for this subtask, they usefully combine with contextual features to yield an F-score of 28.7%, which exceeds the performance of oracle lexical features by 31% relative.

Second, for statements (`s`), contextual features yield no statistically significant difference when added to either "Topo only" or "Pros", while prosodic features offer a small but significant improvement over "Topo only".

Third, for the remaining DA and DA termination types, combining contextual and prosodic features outperforms both feature set types when used alone, and all differences between "Cont" and "Pros" in Table 1 are significant. To compare the contribution of the two, we subtract the F-scores achieved by "Topo only" (which all other systems rely on), and normalize the differences such that unity corresponds to the F-score achieved by "Lex" for each condition. This separates the 9 conditions into two groups, shown in Figure 3. As can be seen in panel (a), DA types signaling intent to retain the floor (`fh` and `h`) and DA types implementing feedback (`b`

[2]We used S. Pado's SIGF implementation of the approximate randomization test (`http://www.nlpado.de/~sebastian/sigf.html`). The data was stratified into intervals corresponding to true DA units, since consecutive 100 ms frames cannot be assumed independent.



(a) Prosody > Context (b) Context > Prosody

Fig. 3. F-score of contextual ("Cont"), prosodic ("Pros"), and combined ("Cont & Pros") systems as a fraction of the F-score achieved by the lexical system ($= 1$), relative to the F-score achieved by the HMM topology alone ($= 0$). Statements (`s`) and interrupted DA termination (*int*) not shown.

and `bk`) are much better detected using prosodic features than using contextual features (in contrast, for both `h` and `b`, the observed c-Opt differences between "Topo only" and "Cont" in Table 1 are not statistically significant). The addition of prosodic features to our "Cont" baseline improves the average F-score over these 4 DA types by 58% relative (and that over all 8 DA types by 28% relative).

Panel (b) of Figure 3 shows the opposite trend for the remaining DA and DA termination types, namely higher F-scores achieved with contextual features than with prosodic features. This trend characterizes either turn beginnings (`fg` and `aa`) or turn ends (`q`, *com*, and *aba*); here, models of the multiparticipant speech/non-speech activity context, indicative of turn construction by interlocutors, were expected to yield good performance. The improvement observed in Table 1 achieved by adding prosodic features to "Cont" is not statistically significant for `fg`, `aa`, or *aba*.

In summary, the combination of contextual and prosodic features delivers improved performance over using either knowledge source alone (except for statements for which contextual features do not help). We note that for several DA types, the performance of the combined system, inclusive of the HMM topology whose effect is normalized out in Figure 3, approaches that of the lexical system[3]. For backchannels, questions, and completed DA punctuation, F-scores reach 87%, 78%, and 92% of their lexical system values, respectively. Furthermore, for holds `h`, the difference between "Cont & Pros" and "Lex" is not statistically significant. Performance for DA termination types is noteworthy in particular, since our systems entertain potential DA boundaries at each 100 ms frame boundary rather than only at word boundaries (which are less frequent), in contrast to work in the literature [1].

There are a number of aspects of the proposed framework which deserve further study. First, the HMM topology used in this work relies on identical sub-topologies for each DA type; pilot experiments indicate that ablating certain states for certain DA types improves performance. Second, aspects of the contextual and the prosodic systems can be optimized both independently, and within the scope of a combined system. This includes the tiling of context in the contextual system and tuning the computation of FFV features [11] to the current task in the prosodic system. Finally, the experiments in this work should be repeated using automatically detected speech activity to determine whether the relative pattern of improvements over DA types holds in a fully automatic setting.

[3]We expect better upper-bound performance by combining all feature types, a task for future work.

5. CONCLUSIONS

We have proposed a framework for text-independent dialog act recognition that combines the multiparticipant speech/non-speech context of our baseline system with features describing the prosody of the target speaker. We find that prosodic features are better than contextual features at recognizing dialog acts deployed to hold the floor or to provide feedback; in these cases, prosody improves the average F-score achieved by our contextual system by 58% relative. In contrast, contextual features outperform prosodic features in recognizing dialog act boundaries, as well as DA types occurring at turn boundaries. The performance of the combined contextual and prosodic system yields F-scores relative to lexical system F-scores which are as high as 78% for questions, 87% for backchannels, 92% for completed-DA boundaries, and 131% for interrupted-DA boundaries. These results suggest that there is significant promise in text-independent features for DA recognition, and possibly for other SLU tasks, particularly when words are not available.

6. REFERENCES

[1] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*, 2005, pp. 1061–1064.

[2] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching DBN," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1303–1314, 2008.

[3] M. Zimmermann, "Joint segmentation and classification of dialog acts using conditional random fields," in *Proc. INTERSPEECH*, 2009, pp. 864–867.

[4] S. Rangarajan, S. Narayanan, and S. Bangalore, "Modeling the intonation of discourse segments for improved online dialog act tagging," in *Proc. ICASSP*, 2009, pp. 5033–5036.

[5] M. Ostendorf, "Transcribing human-directed speech for spoken language processing," in *Proc. INTERSPEECH*, 2009, pp. 21–24.

[6] D. Wyatt, T. Choudhury, and H. Kautz, "Capturing spontaneous conversation and social dynamics: A privacy sensitive data collection effort," in *Proc. ICASSP*, 2007.

[7] S. Parthasarathi et al., "Investigating privacy-sensitive features for speech detection in multiparty conversation," in *Proc. INTERSPEECH*, 2009, pp. 2243–2246.

[8] K. Laskowski and E. Shriberg, "Modeling other talkers for improved dialog act recognition in meetings," in *Proc. INTERSPEECH*, 2009, pp. 2783–2786.

[9] A. Janin et al., "The ICSI Meeting Corpus," in *Proc. ICASSP*, 2003, pp. 364–367.

[10] E. Shriberg et al., "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," in *Proc. SIGdial*, 2004, pp. 97–100.

[11] K. Laskowski, M. Heldner, and J. Edlund, "A general-purpose 32 ms prosodic vector for hidden Markov modeling," in *Proc. INTERSPEECH*, 2009, pp. 724–727.