Modeling Other Talkers for Improved Dialog Act Recognition in Meetings

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

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park CA, USA
[3]International Computer Science Institute, Berkeley CA, USA

kornel@cs.cmu.edu, ees@speech.sri.com

Abstract

Automatic dialog act (DA) modeling has been shown to benefit meeting understanding, but current approaches to DA recognition tend to suffer from a common problem: they under-represent behaviors found at turn edges, during which the "floor" is negotiated among meeting participants. We propose a new approach that takes into account speech from other talkers, relying only on speech/non-speech information from all participants. We find (1) that modeling other participants improves DA detection, even in the absence of other information, (2) that only the single locally most talkative other participant matters, and (3) that 10 seconds provides a sufficiently large local context. Results further show significant performance improvements over a lexical-only system — particularly for the DAs of interest. We conclude that interaction-based modeling at turn edges can be achieved by relatively simple features and should be incorporated for improved meeting understanding.

Index Terms: vocal interaction, cross-speaker modeling, speech/non-speech, dialog acts, meetings

1. Introduction

The automatic understanding of naturally occurring conversations has to date focused largely on techniques relevant to the indexing, retrieval, and summarization of propositional content. For these tasks, automatic segmentation and classification of DAs [1] has been shown to improve performance [2].

However, the focus on propositional content has optimized systems for those intervals of conversation in which such content is deployed, namely away from speaker turn boundaries. Turn initiation and termination are frequently implemented by DA types that account for only a fraction of verbal effort by time. Often for this reason, current DA recognition systems lump many of these DA types together [1]. For example, floor grabbers, which have been shown to correlate with conversational hot spots [4], find themselves in the same group as floor holders, of which there are many more and which exhibit the opposite correlation. As a result, hot spots and other short events related to participant interaction have been under-represented.

In the current work, we investigate what happens at turn edges by considering other participants' speech in the immediate temporal neighborhood. We analyze only the time-aligned speech/non-speech patterns of all participants, a representation that we will refer to as the *vocal interaction* record [5]. Previous work which has explored this representation includes the classification of speaker role [6, 7, 8], the detection of interaction groups [9], the ranking of participants by dominance and influence [10], and the recognition of meeting group actions [12]. However, these applications collect and model the *statistics* describing vocal interaction over long observation intervals. To the best of our knowledge, vocal interaction features for segmenting and classifying talk found in individual utterances have not been modeled or even proposed.

In this research we investigate turn-related-DA recognition in meetings using a novel computational and experimental approach. We ask the following questions:

- To what extent does local interlocutor speech/non-speech activity predict DA type?
- How many interlocutors should be considered?
- How much time around turn edges is needed?
- Can modeling vocal interaction augment the performance of a state-of-the-art lexical DA recognizer?

Our experiments indicate that considering interlocutor speech significantly improves performance; that it suffices to model speech activity inside of a temporal neighborhood of only 10 seconds; that, under our proposed model, only the locally most talkative interlocutor need be considered; and, finally, that our simple speech activity features are complimentary to oracle lexical information, in particular for the detection of those behaviors which tend to occur at turn boundaries.

2. Data

The data used in this work is the ICSI Meeting Corpus, consisting of 75 longitudinal recordings of naturally occurring meetings by several groups at ICSI [13, 14]. We rely on the previously published split of this data into a TRAINSET of 51 meetings, and a DEVSET and a TESTSET of 11 meetings each.

The meetings are provided with lexical forced alignment and DA annotation. We focus on three groups of DA types, the first that of floor mechanisms, including 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 [14], suggesting that lexical content may not adequately distinguish among them. All other speech implements either statements (`s`) or questions (`q`), representing propositional-content DAs. The priors of these 8 DA types by time, for all three datasets, are in the ranges: 1.10–1.18% for `aa`, 2.65–2.86% for `b`, 1.42–1.48% for `bk`, 0.55–0.63% for `fg`, 2.29–3.00% for `fh`, 0.21–0.36% for `h`, 6.53–6.72% for `q`, and 84.83–85.18% for `s`.

3. Methods

As mentioned earlier, inference of DA type in this work is made using only the vocal interaction record of a meeting. This record consists of, for each participant, contiguous "ON" intervals of speech, which we refer to as *talkspurts*, and inter-talkspurt "OFF" gaps. A talkspurt segmentation, comprising the input to our proposed methods, is formed by concatenating the time spans of adjacent human-transcribed words. Because a single talkspurt can implement a sequence of DAs, we treat talkspurts as constructed out of smaller atomic units; for simplicity, we choose these units to be 100 ms frames. The task we consider is the classification of each frame of speech as one of the 8 DA types in Section 2. We propose to decode a meetings one target participant at a time, and to model the speech/non-speech posterior of non-target participants in the target participant's feature space. DA segmentation is not assumed during classification, and boundaries are implicitly postulated where two adjacent frames are assigned dissimilar labels.

We optimize our systems by maximizing the unweighted arithmetic mean over the F-scores for all 8 DA types. Tuning to individual-DA F-scores is part of our final analysis in Subsection 4.3.

3.1. Simple Distance-to-Speech-Edge Features

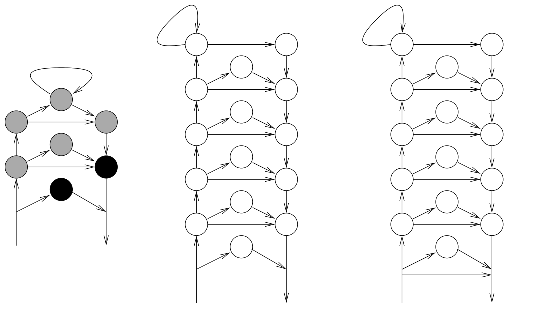
As a preliminary set of context features for a given speech frame at time t, we consider distance to nearby speech from both the target speaker and their interlocutors. For the target speaker, we compute 4 features: the number of frames to the nearest previous speech frame, the number of frames to the nearest next speech frame, and similarly for the nearest non-speech frames. We also compute 3 interlocutor features, namely the number of frames to the nearest previous and nearest next speech frame from *any* non-target participants, as well as the number of concurrent non-target speakers at time t.

Given these features, we train a decision tree using TRAINSET, whose performance on DEVSET for target participant features alone is 16.51%; adding the three non-target participant features improves performance to 18.51%. We note that this approach has two main limitations, namely that (1) frames are assumed conditionally independent, as the decision tree does not leverage sequence information, and (2) the features capture only distance to the nearest edge, and not what lies beyond it.

3.2. A Hidden Markov Model Topology

To address the first limitation, we propose a decoder topology whose elements are shown in Figure 1. A *talkspurt fragment* (TSF) refers to the longest contiguous interval of speech belonging to at most one talkspurt and at most one DA. Each TSF sub-network has a minimum duration constraint of 1 frame, and a maximum no-repeat state sequence of 5 frames to constrain topology size. Since many TSFs are longer than 500 ms, we allow a single self-loop on the center-most state in the TSF sub-network; we prohibit self-loops elsewhere in order to precisely model speech context in the neighborhood of TSF edges.

The network for each DA type consists of one non-DA-terminal and one DA-terminal TSF sub-network; networks for q and s DA types have two additional DA-terminal TSF sub-networks: one for abandoned DAs and one for interrupted DAs. As a result, the complete HMM topology contains 20 distinct TSF sub-networks. Transition from a non-DA-terminal TSF to a DA-terminal TSF must pass through an intra-DA gap (GAP) sub-network; DA networks may be entered through a DA-terminal TSF sub-network for DAs with no intra-DA gaps. Egress states in DA-terminal TSFs are punctuation-bearing. The complete decoder topology involves full connectivity among the 8 DA networks via inter-DA GAPs.



(a) TSF (b) intra-DA GAP (c) inter-DA GAP

Figure 1: Element sub-networks in an HMM topology for conversational speech, with a frame step of 100 ms; states shown in white denote non-speech. In (a), egress states, optionally punctuation-bearing, are shown in black. Note that (b) and (c) differ in that inter-DA GAPs may have zero duration.

Transition probabilities are maximum likelihood estimates inferred from the best Viterbi path through TRAINSET. When only the binary speech/non-speech activity at time t of the target participant is used, this system achieves an average 8-class DEVSET F-score of 20.60%. To include the simple features of the previous section, we model them in log-space using Gaussian mixture models (GMMs). The number of Gaussian components, as well as the emission model weight, are optimized to maximize the average 8-class F-score on DEVSET. The DEVSET performance of this system is 21.93% for target-participant features only, and 24.06% when non-target-participant features are included.

3.3. Improved Vocal Interaction Features

We now turn to the second limitation mentioned in Subsection 3.1, namely that the simple features capture only the nearest talkspurt-edge events; the latter are modeled even when they are temporally very distant. Although we could generalize feature extraction beyond the first edge, we instead propose an alternative approach, relying on only a *local* neighborhood of ΔT seconds around instant t, $t-\Delta T$ to $t+\Delta T$. In our experiments, we have used $\Delta T = 5$ seconds.

We first rank non-target participants, based on their amount of talk in this neighborhood. Vocal interaction features are then extracted for the most talkative non-target participants by considering 0.5 s windows to the left and right of the current instant t, as well as single frames at t. Finally, from each window, we compute the mean of the speech activity posterior from the frames in that window. We denote the feature vector consisting of means drawn from the speech patterns of only the target participant as TARGET. Feature vectors describing the locally most talkative non-target participant, the two locally most talkative non-target participants, and the three locally most talkative non-target participants are denoted OTHER1 $\supset$ OTHER2 $\supset$ OTHER3, respectively. Finally, the concatenation of TARGET and OTHERN will simply be referred to a VOCINTN. An example of interlocutor ranking and VOCINT3 extraction regions is shown in Figure 2.

We model the correlated VOCINTN features using GMMs, following a linear discriminant analysis (LDA) transform. The number of Gaussian components and LDA discriminants, as well as the emission model weight, are separately optimized us-

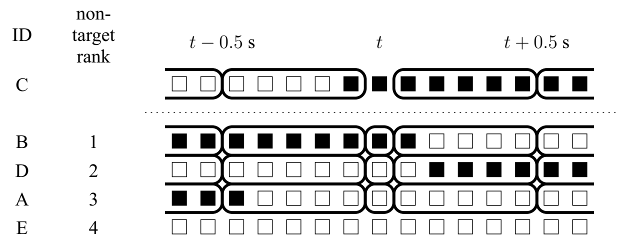


Figure 2: An example of interlocutor rotation and feature extraction at instant t, with time depicted from left to right, when decoding participant C. Non-target participants A, B, D, and E have been ranked according to their amount of talk in the local neighborhood, shown as black squares. Windows for which features are extracted are shown as ovals; a single mean speech activity posterior is computed for each, out to $t - 5.0$ s and $t + 0.5$ s (only windows near t are shown).

ing DevSet. Within the proposed HMM framework, DevSet performance using Target is 24.88%. When Other3 features are included, the performance increases to 28.37%. This represents absolute improvements of 2.95% and 4.31% with respect to the combination of the simple features of Subsection 3.1 and the HMM topology. In the remainder of the work, we rely only on these improved VocInt$N$ features. We note that they offer two additional advantages, to be explored in future work: (1) they rely on a finite context and may be more suitable for online processing; and (2) they are in principle more robust to speech activity insertion and deletion errors.

4. Experiments

This section presents two experiments. In the first, we compare a baseline in which only Target features are modeled to one with Other$N$ features added. In the second, we compare a strong lexical baseline to itself with VocInt3 added.

4.1. Vocal Interaction Features

The average 8-class F-score of the HMM system, using our improved vocal interaction features, is shown in Figure 3 as a function of the size of the neighborhood from which these features are computed. As can be seen, performance for Target and for its combinations with Other$N$ increases sharply as the neighborhood grows to approximately 15 seconds for Target and 10 seconds for the Target+Other$N$ systems.

A second observation is that differences among the Target+Other$N$ systems are quite small, relative to their differences with Target. This indicates that modeling more than just the single locally most talkative non-target participant offers little additional gain. We suspect that the noisy behavior of the Target+Other$N$ curves for larger window sizes is due to our technique of ranking participants during feature extraction; it is possible that participants who speak in close temporal proximity to the current instant t should be ranked higher than those who may speak more but also much sooner or much later.

Over the entire range of explored context sizes, the performance of Target features alone is always lower than the best Target+Other$N$ system by approximately 2-4%. It is possible, given the asymptotic behavior of the Target+Other$N$ curves and the fact that the Target curve is only slowly rising for 40-second contexts, that Target eventually catches up with Target+Other$N$. This requires further investigation. In the meantime, what Figure 3 makes clear is that, at its highest observed performance (achieved using a 40-second context), Target is outperformed by all Target+Other$N$ systems with only 5 seconds of context. This may have implications for real-time conversation processing systems.

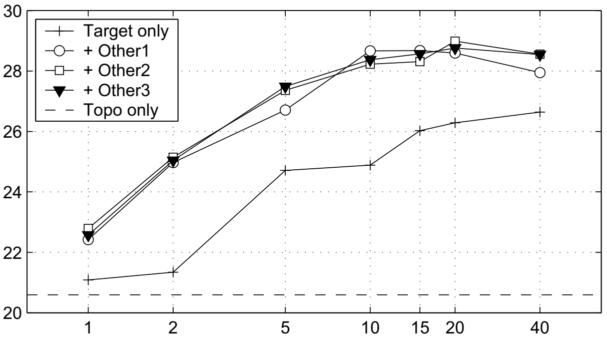


Figure 3: Average 8-class F-score for 5 systems, as a function of the neighborhood size $2\Delta T$, shown on a logarithmic scale in seconds along the x-axis.

4.2. Complimentarity with Lexical Features

In our second experiment, we construct an oracle lexical baseline as the most optimistic single-source measure of performance on this task. We note that DA classification is biased towards lexical information because of the way the data was labeled, making a lexical baseline extremely hard to beat.

The features in our proposed lexical HMM system are the observed lexical bigrams. Each frame of speech occurs closest to the center of a single word; we assign to it the left and right bigrams containing that word. Only those left and right bigrams per state are modeled whose probability of occurrence for at least one DA type exceeds 0.1% in TrainSet. During decoding, emission and transition probabilities are combined using a global weight as for VocInt$N$ systems.

The average 8-class F-score achieved by the lexical HMM system on DevSet is 50.10%, which is indicative of the difficulty of this task, even when perfect word information is present. As a measure of the competitiveness of this system, we have retrained it on the traditional 5-class DA task reported in [1] and re-optimized the lexical model weight. Although its complexity is much smaller than the hidden event language model (HE-LM) in [1], it achieves an error rate of 22.62% on the "lenient" classification error metric (with its own automatic DA segmentation) on the unseen TestSet data, an error rate which is 2.5% lower than that achieved by HE-LM.

To assess feature complimentarity, on the 8-class task, we combine VocInt3 with this lexical baseline (using a second emission model weight). The maximum F-score achieved by the combined system is 52.33%. This represents a 2.23% absolute improvement, and indicates that the proposed VocInt$N$ features offer complimentary information for DA detection.

4.3. Analysis of Performance on Unseen TestSet Data

On TestSet, the average 8-class F-score follows a trend quite similar to that on DevSet. Target improves the performance of the HMM topology alone from 21.81% to 25.48% by 3.67%

DA	OTHER3 added to TARGET		VOCINT3 added to LEX	
	abs,%	rel,%	abs,%	rel,%
fg	10.4 → 13.7	+31.8*	24.5 → 27.0	+9.8*
h	1.1 → 6.3	+485.6*†	41.5 → 42.3	+2.0*
fh	21.7 → 25.6	+18.3*†	63.5 → 64.5	+1.5
b	56.7 → 57.8	+1.9*†	77.0 → 77.9	+1.1*
bk	12.6 → 14.9	+18.5*	56.3 → 56.0	−0.5
aa	8.7 → 13.0	+49.4*†	40.0 → 42.0	+5.0*†
q	23.4 → 26.3	+12.3*†	39.8 → 42.5	+6.8*†
s	91.4 → 91.3	−0.08*	93.3 → 93.5	+0.2*†
int	10.7 → 22.6	+111.3*†	21.9 → 34.1	+56.0*†
aba	7.0 → 6.6	−6.1	13.0 → 14.4	+10.3 †
ter	61.4 → 62.1	+1.2*†	69.1 → 69.6	+0.7 †

Table 1: Absolute ("abs") and relative ("rel") improvements in F-score on EVALSET, for individual DA conditions, obtained by including non-target participant featurees in the non-lexical baseline, and those obtained by including all VOCINT3 features in the lexical baseline. "*" indicates significance at $p < 0.01$ using a randomization test; "†" indicates significance at $p < 0.05$ when labels are stratified into talkspurts; *int* = interruption, *aba* = abandonment, *ter* = (normal) termination.

absolute. Inclusion of OTHER3 improves it further to 29.33%, by 3.85% absolute. When included in the lexical baseline, which achieves 52.98% on TESTSET, VOCINT3 improves performance to 54.74% by 1.76%.

We present EVALSET F-scores for systems optimized for specific DA types in Table 1; optimization consisted of tuning model combination weights only. The table also shows F-scores for the retrieval of frames implementing specific DA termination types; these scores were not included in the average 8-class F-score during development.

It can be seen that the relative improvement in F-score for most DA types is high, when TARGET features are augmented with OTHER3 features; this is due in large part to their rather poor absolute F-scores. All positive improvements were shown to be statistically significant at the frame level, at $p < 0.01$, using a randomization test[1]. Performance using lexical features is much higher, for all DA and DA-termination types. When the VOCINT3 model is used alongside the oracle lexical model, the improvements observed in the left part of the table become smaller. Nevertheless, for all DA types but acknowledgements (bk), the relative improvements are positive. Floor grabbers (fg) exhibit the largest relative increase, followed by questions (q), which tend to terminate speaker turns, and asserts (aa). The largest improvement is seen for the interruption of the ongoing DA (of 56%), which is a single-frame event. Many of the F-score increases are significant not only at the frame level, but also when label sequences are stratified into talkspurts. Finally, we note that significant but small improvements are also observed for statements, backchannels, and normal DA termination events.

5. Conclusions

We have defined a new set of features for DA recognition in multi-party meetings, to aid in the detection of phenomena occurring at speaker turn edges. The features describe the local distribution of speech/non-speech activity for the target participant and for his/her interlocutors. Our results indicate that inclusion of interlocutor behavior significantly improves overall DA recognition, with the largest gains for the short, turn-related DAs under-represented in standard systems. We find that only the single locally most talkative interlocutor matters, corroborating findings in conversation analysis regarding the orderliness of floor change and its immutability to group size [3]. A temporal context of 5 seconds in each direction appears sufficient for capturing inter-participant DA dependency. We also demonstrate that when combined with a purposely optimistic oracle lexical system, the proposed features yield consistent improvements in average 8-class F-score. Most importantly, we observe large relative gains for specifically those behaviors that initiate turns (floor grabbers, of 9.8%), and those which accompany their completion (questions, of 6.8%, and interruption, of 56%). Together, these results indicate that the modeling of interaction at turn edges can be achieved with surprisingly simple-to-compute features. Because the features and modeling approach are text-independent, they offer the possibility of general applicability across different languages and speaking styles.

6. Acknowledgements

We would like to thank an anonymous reviewer who provided helpful comments on structure and argumentation.

7. References

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

[2] Kathol, A. and Tur, G., 'Extracting question/answer pairs in multiparty meetings", in Proc. ICASSP, 5053-5056, 2008.

[3] Sacks, H., Schegloff, E. and Jefferson, G., "A simplest semantics for the organization of the turn-taking in conversation", in *Language*:**50**, 696-735, 1974.

[4] Wrede, B. and Shriberg, E., "The relationship between dialogue acts and hot spots in meetings", in Proc. ASRU, 180-185, 2003.

[5] Dabbs, J. and Ruback, R., "Dimensions of group process: Amount and structure of vocal interaction", in *Advances in Experimental Psychology*:**20**, 123-169, 1987.

[6] Banerjee, S. and Rudnicky, A., "Using simple speech based features to detect the state of a meeting and the roles of the meeting participants", in Proc. INTERSPEECH, 2189-2192, 2004.

[7] Laskowski, K., Ostendorf, M. and Schultz, T., "Modeling vocal interaction for text-independent participant characterization in multi-party conversation", in Proc. SIGdial, 148-155, 2008.

[8] Favre, S., Salamin, H., Dines, J. and Vinciarelli, A., "Role recognition in multiparty recordings using social affiliation networks and discrete distributions", in Proc. ICMI, 29-36, 2008.

[9] Brdiczka, O., Maisonnasse, J. and Reignier, P., "Automatic detection of interaction groups", in Proc. ICMI, 2005.

[10] Rienks, R., Zhang, D., Gatica-Perez, D. and Post, W., "Detection and application of influence rankings in small-group meetings", in Proc. ICMI, 2006.

[12] McCowan, I., Bengio, S., Gatica-Perez, D., Lathout, G., Barnard, M. and Zhang, D., "Automatic analysis of multimodal group actions in meetings", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*:**27**(3), 305-317, 2005.

[13] Janin, A. et al, "The ICSI Meeting Corpus", in Proc. ICASSP, 364-367, 2003.

[14] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J. and Carvey, H., "The ICSI MRDA Corpus", in Proc. SIGdial, 97-100, 2004.

[1] We used Sebastian Pado's SIGF impementation, found at http://www.nlpado.de/˜sebastian/sigf.html.