A Frame-Synchronous Prosodic Decoder for Text-Independent Dialog Act Recognition

Kornel Laskowski

Language Technologies Institute
Carnegie Mellon University
Pittsburgh PA, USA
kornel@cs.cmu.edu

Abstract

Dialog act (DA) recognition is an important intermediate task is speech understanding systems. Although past research has demonstrated that prosody can improve the performance of recognizers relying primarily on words, how prosody fares on its own is not well understood. The current work continues an ongoing investigation into settings in which both words and word boundaries are unavailable, whether for privacy, security, speed, or availability of technology reasons. A system is presented with long acoustic frames, which renders the modeling of prosodic context tractable. The system is then extended by concatenating features computed for temporally proximate frames, from both the target speaker and from non-target interlocutors. Experiments indicate that the increased frame size and target-speaker prosodic context improve recognition performance, in particular for floor holders, accepts, and DA termination types. Non-target-speaker prosodic context is shown to have a large positive impact on the detection of DA interruption. These results suggest that the improved framework holds promise for the general decoding of prosodic phenomena in spontaneous speech, independently of speech recognition.

Index Terms: prosodic features, dialog act tagging, cross-speaker modeling, HMM modeling of prosody, meetings

1. Introduction

Dialog act (DA) recognition, or the simultaneous segmentation and classification of conversational speech into dialog acts, is known to be important in speech understanding systems. Examples of useful DA distinctions are those between statements, questions, acknowledgments, and floor control mechanisms. Although past work has focused largely on lexical information, prosodic features have been shown to be beneficial and complimentary to words, for segmentation [1, 2, 3, 5] as well as for classification [1, 3, 4, 5, 6]. However, the computation of most prosodic features makes some use of word information, at least in the form of word boundaries. To the growing community of researchers interested in privacy-sensitive settings, where words are not available because spectral features cannot be computed [7, 8], such systems offer only a limited starting point.

Recently, a text-independent DA recognizer, based on a hidden Markov model (HMM) decoder, was proposed to compare the performance of multiparticipant speech/non-speech context features to that of prosodic features [9]. The decoder operated at a frame step of 100 ms, a size commensurate with the durations of dialogue phenomena which are of interest here, but for expediency relied on an acoustic feature vector which was computed every 8 ms, using frames 32 ms in duration [10]. This required a run-time frame-rate conversion [9].

There are several reasons why this may be suboptimal. First, the perceptual quantities relevant to HMM-based DA recognition may be better captured with frame-level features computed for timespans longer than 32 ms. However, it is not known whether the prosodic features proposed in [10], which model intonation but avoid the direct inference of pitch, are useful when computed for longer frames. Second, parity with the decoder frame step during feature computation potentially entails order-of-magnitude savings in time complexity, making it a valid pursuit in its own right. Finally, and most importantly, a longer feature computation frame step would render tractable the explicit modeling of long-term prosodic interdependencies. Of course, extending the feature computation frame step by a factor of 12.5 also reduces by the same factor the number of feature vectors available for model training. For rarely occurring phenomena, this may undermine the above potential benefits.

The current work addresses these concerns by explicitly asking the following three questions:

1. Can the instantaneous feature vector of [9, 10] be extended to describe audio frames whose size is larger than 32 ms, at a frame step of 100 ms, without negative impact on DA recognition performance?
2. Does feature-space combination with temporally adjacent features from the target participant improve DA recognition?
3. Does feature-space combination with temporally adjacent features from non-target participants improve DA recognition?

Beginning with the decoder of [9] as a baseline, described in Section 3, experiments are presented which answer the above questions, in Section 4. It is shown that extending the feature computation frame step and frame size of the baseline decoder, by redefining the feature extraction algorithm, leads to improved DA recognition. The subsequently possible modeling of temporally proximate prosody from the target speaker is shown to improve both segmentation and classification performance. Modeling non-target-speaker prosody appears to further improve DA segmentation, at a small cost to classification accuracy. These findings suggest that the proposed framework offers a promising avenue for the inference of prosodic phrasing, whether in parallel with speech recognition or on its own.

2. Data

The data used in this work is the ICSI Meeting Corpus, consisting of recordings of 75 naturally occurring meetings held by

several groups at ICSI [11, 12]. The previously published split of this data, into a TRAINSET of 51 meetings, and a DEVSET and a TESTSET of 11 meetings each, is retained throughout.

The corpus is provided with lexical forced alignment and DA annotation. Three groups of DA types are of interest, 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`), as well as accepts (`aa`). These six types have been reported to share a common vocabulary [12], 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`. DAs are terminated in three mutually exclusive ways: they are completed, interrupted, or abandoned.

3. Baseline

In a word-synchronous setting, the role of a DA recognizer is to assign to each word exactly one DA type, and to potentially assign to each word's end exactly one DA boundary type. When words and word boundaries are not available, however, a different underlying unit must be selected. In the current work, as in [9, 13], that unit is the 100 ms frame (of speech). The role of a frame-synchronous DA recognizer, then, is to assign *to each such frame* exactly one DA type and to potentially assign *to each such frame's end* exactly one DA boundary type. For the experiments presented here, the speech/non-speech segmentation is taken from forced-alignment-mediated human reference transcription [12].

3.1. Topology & Transition Probabilities

Since contiguous intervals of speech, or *talkspurts*, may implement a sequence of DAs, the recognizer must split talkspurts, as well as merge them across non-speech intervals where necessary. To facilitate discussion, the term *talkspurt fragment* (TSF) is used here to denote a contiguous interval of speech belonging to exactly one talkspurt and to exactly one DA.

The HMM topology in the present work is that proposed in [13]. Each DA type is represented by an identical subtopology, with states modeling speech at the beginning, middle, and end of DA-non-terminal and DA-terminal TSFs, and non-speech both within and between DAs. Final DA-terminal TSF states bear one of the three possible DA boundaries. DA type subtopologies are fully connected, allowing any DA type to follow any other.

The state-to-state transition probabilities are trained using the best Viterbi forced alignment path of the annotated TRAINSET. During decoding, transition probabilities are combined with emission probabilities additively in the log domain, using a mixing weight λ with $\log_2 \lambda \in \{-4, -3, \ldots, +7, +8\}$. For each experiment, λ is tuned to maximize the mean F-score over the 8 DA types over DEVSET, leading to a globally optimized system subsequently referred to as "g-Opt".

3.2. Prosodic Features

Prosodic observables are characterized by the feature vector proposed in [10], containing correlates of loudness, voice quality, speech rate, and variation in pitch. These are the frame energy and the first-order difference in energy, the normalized maximum of the first peak in the auto-correlation spectrum, the cosine distance between the Mel spectra, and between the log-Mel spectra, of the left and right half of each frame, and the 7-element fundamental frequency variation (FFV) spectrum representation, respectively.

As in [9, 10], the feature vector is computed for frames 32 ms in size, at a frame step of 8 ms. The 32 ms constraint is imposed by the current definition of the FFV spectrum, as briefly explained here. The spectrum is a one-dimensional function of a single parameter, the *frequency dilation factor* ρ, which is expressed in octaves per second. The spectrum indicates, for each value of ρ, the similarity between the magnitude frequency spectrum of the left half of an audio frame and a ρ-dilated version of the magnitude frequency spectrum of the right half of the same frame. The location ρ^* of the maximum in the spectrum, for a voiced frame, indicates the instantaneous change in fundamental frequency; rather than finding this maximum, the entire spectrum is passed through a filterbank. The filterbank's 7 filters include 5 filters capturing various rates of positive and negative change in F0, plus 2 filters which serve a normalization purpose.

3.3. Prosodic Emission Probabilities

The 12-element feature vector is modeled by first rotating it using linear discriminant analysis; the number of retained discriminants is set to maximize the mean F-score on DEVSET. Each state in the topology has its own associated emission probability Gaussian mixture model; the number of Gaussian components is set to optimize the same criterion, and is identical for all states. Because the decoder runs at a frame step of 100 ms, but the feature vectors in the baseline are computed at a frame step of 8 ms, the emission score used during Viterbi search is the average log-likelihood of 12.5 consecutive feature vectors.

3.4. Performance Characterization

The unavailability of words eliminates the possibility of scoring systems using traditional, word-based metrics [1, 14]. The primary metric selected to develop the presented systems is the unweighted mean of the frame-level F-scores for the 8 DA types of interest [13]. Maximizing this metric leads to a single, "globally optimized" ("g-Opt") system. To understand how features affect specific DA types, "condition-optimized" ("c-Opt") systems are constructed as in [9] by varying λ to optimize the F-score for specific DA types; "c-Opt" systems for other metrics, such as the overall classification error, are selected in the same way. In addition, F-scores are computed for specific DA termination types, as well as for "termination in general" by collapsing the three different DA termination types into one. Also computed is the NIST segmentation error, but this metric must be interpreted with caution when comparing against existing lexical systems. There, potential DA boundaries are entertained only at word boundaries [1], which are unavailable to the systems in the current work; as a result, the number of potential false alarms is much higher here, as DA boundaries are entertained at every speech frame.

The performance of the baseline systems is shown in columns 3 and 4 of Table 1, along with the performance of the HMM topology alone in column 2 (these numbers are taken from [9]). As can be seen, the baseline offers much improved performance over the topology-only system on all metrics, except on the F-score for statements and the NIST segmentation error using the "g-Opt" system. For holds and accepts, the "g-Opt" system outperforms the two "c-Opt" systems optimized

	Topo only	Baseline 32 ms 8ms (12.5 frames)		Experiment 1 256 ms 100 ms (1 frame)		Experiment 2 256 ms 100 ms (11 frames)		Rel. Improvement in %, Baseline to Experiment 2	
	%	g-Opt %	c-Opt %	g-Opt %	c-Opt %	g-Opt %	c-Opt %	g-Opt %	c-Opt %, sig
DA Types									
mean F	21.8	31.5	33.7	35.5	36.2	36.6	37.9	+16.2	+12.5 —
F, floor holder `fh`	11.3	37.7	39.5	45.2	45.2	45.8	48.2	+21.5	+22.0 **
F, hold `h`	0.0	25.0	17.1	25.3	20.6	21.6	21.6	−13.6	+26.3 *
F, floor grabber `fg`	0.0	7.2	7.2	8.2	8.2	10.1	10.1	+40.3	+40.3 *
F, backchannel `b`	57.1	48.0	64.6	57.5	64.4	59.9	64.2	+24.8	−0.6
F, acknowledgment `bk`	3.2	19.0	20.9	25.2	25.2	25.3	25.3	+33.2	+21.1 **
F, accept `aa`	2.6	9.5	8.9	17.5	17.5	21.9	22.5	+130.5	+152.8 **
F, statement `s`	91.4	85.8	91.8	88.8	91.9	88.4	92.0	+3.0	+0.2 **
F, question `q`	8.8	19.6	19.6	16.6	16.6	19.4	19.4	−1.0	−1.0
classification error	17.0	25.9	16.6	21.0	15.9	21.7	15.8	−16.2	−4.8 —
DA Termination Types									
F, completed	53.1	59.1	59.1	59.9	59.9	62.7	63.8	+6.1	+8.0 **
F, interrupted	0.0	10.5	11.8	6.7	9.6	14.6	14.6	+39.1	+23.7
F, abandoned	0.0	2.4	3.6	2.4	4.3	6.1	7.0	+154.2	+94.4 **
any type, F	53.9	62.6	62.6	62.4	62.4	66.4	66.4	+6.1	+6.1 —
NIST error	64.7	66.5	63.0	66.5	63.0	66.1	58.5	−0.6	−7.0 —

Table 1: EVALSET performance for the topology alone, the baseline systems of Section 3, the systems in Experiment 1 (Section 4.1), the systems in Experiment 2 (Section 4.2), and the relative improvement from the baseline systems to the systems in Experiment 2. The last column includes assessment of statistical significance of the difference in F score in the "c-Opt" condition, at $p < 0.005$, using an approximate randomization test. "**" indicates significance when labels are stratified using reference DA boundaries, "*" indicates significance only for unstratified labels; rows marked with "—" were not tested.

specifically for the detection of these two DA types. This appears to be an example of overfitting to DEVSET.

4. Experiments

4.1. Increasing the Frame Step to 100 ms

In a first suite of experiments, the feature computation frame step of 8 ms is extended to the decoder frame step of 100 ms. To avoid missing audio, the frame size must also be increased; experimentation with DEVSET led to a size of 256 ms. While this presents no problem for those features which correlate with loudness, voice quality, and speaking rate, the computation of FFV features at this frame size requires some consideration.

In increasing the frame size along a continuum, Hann profiles were adopted for the outside as well as the inside edges of the left and right window functions. Parameters governing window extent[1] were simply scaled by a factor of 8; no DEVSET improvement was observed with further perturbation. Adding two additional filters to the filterbank, representing very quickly decreasing and very quickly increasing fundamental frequency, led to a small improvement in mean F-score. This yielded an FFV feature vector of 9 coefficients, extending the complete feature vector from 12 to 14 features.

The results for EVALSET, unseen during these developments, are shown in columns 5 and 6 of Table 1. The single system whose mean F-score was maximized ("g-Opt") shows a 4.0% absolute improvement in mean F-score over the baseline; a smaller improvement of 2.5%abs is observed in the mean computed over the 8 DA-specific "c-Opt" systems. These improvements accompany reduction in the classification error.

[1]Namely t_{sep}, t_{int}, and t_{ext}; cf. [10, 15].

The largest "c-Opt" improvements are observed for floor holders, acknowledgments, and accepts (as well as for backchannels for the "g-Opt" system), and the largest drop in F-score for questions. The latter is likely due to the many different types of questions in meeting data, with corresponding variability in prosodic form [6]. (Holds continue to be better detected by the "g-Opt" system than by the "c-Opt" system optimized for their detection, indicating the same overfitting to DEVSET as was observed for the baseline.) In general, it can be concluded that, despite training data ablation, redefinition of FFV computation does not erode overall DA recognition.

4.2. Temporally Proximate Target-Speaker Frames

In a second experimental suite, the impact of extending the feature vector for each frame with features computed from the previous 5 frames and the subsequent 5 frames is explored. This provides a context of 1 second around the frame of interest, where an inflection point was observed in preliminary work. The results are shown in columns 7 and 8 of Table 1. A mean F-score increase is observed over the results of Experiment 1 (however, in the "g-Opt" condition, the 1.1% increase in F-score is accompanied by a 0.7% increase in classification error). Improvements are largest in absolute terms, given both "g-Opt" and "c-Opt" systems, for floor holders and accepts, a trend that is also observed for DEVSET (not shown).

Columns 9 and 10 show the relative improvements of Experiment 2 systems with respect to the baseline. In addition to consistent improvements for floor holders and accepts, and some recovery in the detection of questions, DA terminations are better detected, for both completed and abandoned DAs. This improvement is somewhat serendipitous, as the mean F-

score towards which system components were originally optimized ignored DA terminations.

4.3. Temporally Proximate Non-Target-Speaker Frames

The third and final set of experiments explores what happens when one includes in the feature vector of each frame not only the target speaker's adjacent features, but also the features from non-target speakers. In [13], it was shown that when modeling non-target-speaker speech/non-speech activity for this task, it was sufficient to include the speech/non-speech activity from only the locally most talkative other speaker. That idea is reemployed here, using the 1-second temporal context of reference speech/non-speech activity around each frame to identify that non-target speaker and to include their 11 feature vectors in the feature vector for the current frame. The features are computed and modeled for these context frames regardless of whether they are speech or non-speech frames. Ties when ranking non-target speakers are resolved by random selection.

The results are presented in Table 2; columns 7 and 8 from Table 1 are included for ease of comparison. As can be seen, the mean F-score over the 8 DA types decreases when non-target speaker features are included, by $< 1.5\%$, but including them appears to improve the detection of whether a DA is interrupted: the F-score nearly doubles from 14.6% to 28.4%. Although the F-scores for normally completed and abandoned DA termination types fall slightly, this gain yields an increase in the F-score for termination in general ("any type" in the tables) by 1.8% in the "c-Opt" condition, as well as a a 2.5% reduction in the "lenient" NIST boundary error.

	Experiment 2		Experiment 3	
	g-Opt	c-Opt	g-Opt	c-Opt
DA Types				
mean F	36.6	37.9	35.8	36.4
classification error	21.7	15.8	21.4	16.1
DA Term. Types				
F, completed	62.7	63.8	61.0	63.4
F, interrupted	14.6	14.6	28.4	28.4
F, abandoned	6.1	7.0	4.6	5.4
any type, F	66.4	66.4	66.9	68.2
NIST error	66.1	58.5	65.9	56.0

Table 2: EVALSET performance for Experiments 2 and 3.

5. Conclusions

This work has described improvements to a novel frame-synchronous HMM-based decoder of prosodic phenomena. The modifications, which were three in number, were assessed on the task of simultaneous segmentation and classification of spontaneous multi-party speech into dialog acts. First, the computation of fundamental frequency variation features was extended to acoustic frames 256 ms in duration. This was observed to have positive impact on overall DA recognition, despite the entailed ablation of training material by an order of magnitude. It was also argued that this modification both reduces time complexity and facilitates the modeling of longer prosodic context. The latter claim was explored by concatenating prosodic features from temporally adjacent frames, within a 1-second context window. Such inclusion of the target speaker's features was shown to yield significant improvements in DA recognition, particularly for floor holders, accepts, and DA boundaries. Inclusion of features from the locally most-talkative non-target speaker was shown to be slightly deleterious for classification of DA type, but to improve the detection of interrupted statements and questions. These results are encouraging, and recommend further development of DA recognizers which are both text-independent, relying on no word or word boundary information, and privacy-sensitive, making use of no features from which speaker identity can be inferred. The proposed framework offers broad scope for that development, in both offline and online applications.

6. Acknowledgments

The suggestion to compute FFV features using longer frames is due to Mari Ostendorf. The author would also like to thank an uncommonly meticulous reviewer whose comments had significant impact on the final version of this manuscript.

7. References

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

[2] Kolář, J., Shriberg, E., and Liu, Y., "Using prosody for automatic sentence segmentation of multi-party meetings", Proc. TSD, 629–636, Brno, Czech Republic, 2006.

[3] Dielmann, A. and Renals, S., "Recognition of dialogue acts in multiparty meetings using a switching DBN", in IEEE Trans. Audio, Speech, and Language Proc., 16(7):1303–1314, 2008.

[4] Rangarajan, S. V. K. et al, "Modeling the intonation of discourse segments for improved online dialog act tagging", in Proc. ICASSP, 5033-5036, Taipei, Taiwan, 2009.

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

[6] Boakye, K., Favre, B., and Hakkani-Tür, D., "Any questions? Automatic question detection in meetings", in Proc. ASRU, Merano, Italy, 2009.

[7] Wyatt, D., Choudhury, T., and Kautz, H., "Capturing spontaneous conversation and social dynamics: A privacy sensitive data collection effort", in Proc. ICASSP, 213–216, Honolulu HI, USA, 2007.

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

[9] Laskowski, K. and Shriberg, E., "Comparing the contributions of context and prosody in text-independent dialog act recognition", in Proc. ICASSP, Dallas TX, USA, 2010.

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

[11] Janin, A. et al, "The ICSI Meeting Corpus", in Proc. ICASSP, 364–367, Hong Kong, China, 2003.

[12] Shriberg, E. et al, "The ICSI MRDA Corpus", in Proc. SIGdial, 97–100, Cambridge MA, USA, 2004.

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

[14] Liu, Y. and Shriberg, E., "Comparing evaluation metrics for sentence boundary detection", in Proc. ICASSP, 185–189, Honolulu HI, USA, 2007.

[15] Laskowski, K. and Edlund, J., "A Snack Implementation and Tck/Tk Interface to the Fundamental Frequency Variation Spectrum Algorithm", in Proc. LREC, La Valletta, Malta, 2010.