On Speaker-Specific Prosodic Models for Automatic Dialog Act Segmentation of Multi-Party Meetings

Jáchym Kolář[1,2], *Elizabeth Shriberg*[1,3], *Yang Liu*[1,4]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic
[3]SRI International, Menlo Park, CA, USA [4]University of Texas at Dallas, TX, USA
{jachym,ees,yangl}@icsi.berkeley.edu

Abstract

We explore speaker-specific prosodic modeling for dialog act segmentation of speech from the ICSI Meeting Corpus. We ask whether features beyond pauses help individual speakers, and whether some speakers benefit from prosody models trained on only their speech. We find positive results for both questions, although the second is more complex. Feature analysis reveals that duration is the most used feature type, followed by pause and pitch features. Results also suggest a difference between native and nonnative speakers in feature usage patterns. We conclude that features beyond pauses are useful for dialog act segmentation in natural conversation, and that for some speakers, speaker-specific training yields further gains.

Index Terms: prosody, dialog act segmentation, meetings.

1. Introduction

An area of growing interest in the spoken language technology community is the automatic processing of multi-party meetings. Important tasks in this domain include automatic meeting browsing, summarization, information extraction and retrieval, and machine translation [1, 2]. Approaches to these tasks are typically based on natural language processing techniques that are trained on formatted input, such as text. But the output of a meeting recognizer is an unstructured stream of words. The goal of this paper is to segment the speech from each talker into meaningful units such as sentences or, as in our case, dialog acts (DAs). Unlike previous work, we explore both features beyond pauses, and speaker-specific modeling of prosody for this task.

We define the segmentation task as a two-way classification problem, in which we must label each inter-word boundary, as either a within-unit boundary, or a boundary between DAs. For example, in the utterance "yes we should be done by noon", there are two dialog acts: "yes" (an answer), and "we should be done by noon" (a statement). Each ends in a segmentation boundary.

Previous efforts in sentence and DA segmentation have studied the role of both lexical and prosodic features for data from news broadcasts and from spontaneous telephone conversations [3, 4, 5, 6, 7, 8]. Work on multi-party meetings has been more recent, and has generally examined the use of prosody for segmentation using only pause information [9, 10, 11]. An exception is [12], which showed overall improvements to a speaker-independent prosody model by using prosodic features beyond pauses.

In this paper we take a closer look at prosodic modeling for segmentation by looking at individual speakers. In many meeting applications, the speaker is known and recorded on a separate channel. This presents the opportunity for adapting models to the individual talker. Speaker adaptation in the cepstral domain is widely used in automatic speech recognition, but much less is known about speaker-specific variation in prosodic patterns, beyond basic f_0 normalization. Studies in speech synthesis and automatic speaker recognition have used prosodic variation successfully, but to our knowledge modeling stylistic prosodic variability for sentence boundary recognition has to date been mentioned only anecdotally in the literature [3, 13, 14].

We ask two main questions about speaker variation in prosodic marking of sentence boundaries, using 20 speakers from the ICSI meeting corpus. First, we ask whether individual speakers benefit from modeling more than simply pause information. Second, we explore whether speakers differ enough from an overall (speaker-independent) model of prosody to benefit from a model trained on only their speech. Given the much smaller amount of data available for training speaker-specific models, we expect that most talkers are best described by the larger speaker-independent model. If some speakers do show a win from speaker-dependent modeling, however, despite the much smaller amount of data, this would suggest interesting areas for further research on prosodic adaptation.

2. Method

2.1. Data and experimental setup

The ICSI meeting corpus [15] contains approximately 72 hours of multichannel conversational speech data and associated human transcripts. This corpus was manually annotated for DAs [16].

In order to focus on the aspect of the speaker-dependent prosodic characteristics, and to use the DA boundaries marked by human labelers, we use forced alignment of human transcripts in this work. Similarly, to avoid confounds with independent issues in far-field speech recognition, we used audio from close-talking microphones.

We selected the top 20 speakers in terms of total words. The resulting set contains 17 males and 3 females; 12 speakers are native English speakers and 8 are nonnative speakers. Each speaker's data was split into a training set ($\sim$70% of data) and a test set ($\sim$30%), with the caveat that a speaker's recording in any particular meeting appeared in only one of the sets. Because of data sparseness, we did not use a separate development set, but rather jackknifed the test set (one half of the test data was used for tuning weights for the second half, and vice versa). The total training set for speaker-independent models (comprising training portions of

the 20 analyzed speakers, as well as all data from 62 other less-frequent speakers) contained 567k words. Data set sizes for individual speakers are shown in Table 1. We use the official corpus speaker IDs. The first letter of the ID denotes the sex of the speaker ("f" or "m"); the second letter indicates whether the speaker is a native ("e") or nonnative ("n") speaker of English.

2.2. Prosodic features

We originally developed a database of 270 prosodic features (inspired by [3, 17]) that capture pause, pitch, duration, and energy information associated with each word boundary. Features were extracted directly from the automatically aligned speech signal. After initial experiments, we chose a subset of 32 features that proved useful in related work on the same data.

Pause features consist of the pause duration after the current, the previous, and the following word. Duration features include phone-normalized durations of vowels, final rhymes, and words; normalization statistics were generated from the entire database. We did not use raw duration features, since although they aid performance, they correlate with lexical features that should be modeled in a language model. Certain frequent DAs (esp. backchannels) have small set of words, so raw durations may capture those words rather than prosody. Pitch features include the minimum, maximum, and mean values of f_0, f_0 slopes, and the differences and ratios of values across word boundaries. Pitch features are extracted from both the raw f_0 values and from an f_0 contour stylized by a piece-wise linear function. Energy features were represented by the maximal, minimal, and mean frame-level RMS values, using both raw and per-channel normalized values.

2.3. Classifiers

As in past work on segmentation, we used decision tree classifiers as a prosody model, since they handle features with undefined values, are easy to interpret, and yield good results. However, since DA boundaries occur only at approximately 16% of all the word boundaries in our corpus, we need a way to cope with the problem of data skew. One solution is to train classifiers on data downsampled to equal class priors. To take advantage of all available data, we apply ensemble sampling instead of simple downsampling. Ensemble sampling is performed by randomly splitting the majority class into int(N) nonoverlapping subsets, where N is the ratio between the number of samples in the majority and the minority classes. Each subset is then joined with all the minority class samples to form int(N) balanced sets to train classifiers. It is also advantageous to employ bagging [18], which decreases classifier variance by averaging results obtained by multiple classifiers. For bagging, multiple classifiers are trained from different datasets sampled with replacement from the original training set. We used a combination of these two methods, referred to as *ensemble bagging* [19]. When applying the classifiers on (the imbalanced) test data, we adjust the resulting posteriors to take into account the original class priors.

3. Experimental results and discussion

We measure model performance using a "boundary error rate" [3]:

$$E = \frac{I + M}{N_W} \quad [\%] \tag{1}$$

where I denotes the number of false DA boundary insertions, M the number of misses, and N_W the number of words in the test set.

3.1. Pause-only vs. richer set of prosodic features

The first problem we investigate is whether there is any gain from using a richer set of prosodic features rather than pause information alone, for a speaker-independent model as applied to specific speakers. Table 1 shows the results using two different prosodic feature sets (pause only versus all prosodic features) for each speaker. The speakers displayed in the table are sorted according to the total number of words they have in the corpus. As shown, the richer prosodic feature set (SI-All) yields a significantly better performance than the pause-only model (SI-Pau), for 19 of the 20 speakers. The relative error rate reduction is also provided, indicating that differences across speakers on this measure, interestingly, do not appear to be correlated with the amount of training data. They may thus reflect differences in speaking styles, although other factors such as robustness of feature extraction or production of different rates of DA types, may also play a role.

Because of space limitations, we present detailed results for only the speaker-independent models. Using speaker-specific models, the rich prosodic feature sets achieved better performance for 17 speakers, while for 3 speakers (me003, me025, and fn002), the pause-only model performed better. The particular boundary error rates may be found in the column "SD-Pau" in Table 1 and in the column "SD" in Table 2, respectively.

3.2. Speaker-independent vs. speaker-dependent models

The second problem we investigate is whether some speakers may benefit from speaker-dependent training, despite significantly less data for SD than for SI models. Table 2 compares performance of SI, SD, and interpolated (SI+SD) models using the rich prosodic feature sets. We interpolate posterior probabilities of the two models using

$$P(X) = \lambda P_{SI}(X) + (1 - \lambda) P_{SD}(X) \tag{2}$$

where $P_{SI}(X)$ denotes the speaker-independent and $P_{SD}(X)$ speaker-dependent posterior, and λ is a weighting factor estimated using the jackknife approach as described in Section 2.1.

We also present chance performance – the error rate achieved by classifying every word boundary into the class with the highest prior probability (which is "no DA boundary" in our case). Chance performance reflects a speaker's relative rate of various DA types. For instance, a high chance error rate typically correlates with a speaker who produces many short DAs, such as backchannels. To enable comparison of results across speakers we report the relative error reduction (RER) with respect to chance error for the best model of each speaker. We also show the percentage of the data used for training SD models relative to the data available for training the SI model (DP).

Results in Table 2 indicate that the SD model is better than the SI model for 4 of the 10 most frequent speakers, and for 5 of the 20 speakers. The SD or SI+SD model is better than the SI model for 5 of 10 and for 7 of 20 speakers. We used a Sign test for statistical significance measurement. Four speakers (me011, mn007, fe016, mn005) had results significant at $p <= .05$ or better; one speaker (fn002) was significant at $p <= .10$. Although only some speakers show these improvements (while many others show poor results from SD modeling), the finding is important. If a speaker shows significantly improved results using a model trained on far less data than the SI model, this suggests that the speaker's prosodic marking of DA boundaries differs from that of

Table 1: *DA classification performance (error rate in %) of various models for each speaker, along with the data set size. ID=speaker ID, # Train and # Test denote the number of words in the training and test sets for each speaker, SI-Pau and SD-Pau indicate speaker-independent and speaker-dependent models using only pause information, SI-All is the speaker-independent model using all the prosodic features, and RER denotes relative boundary error rate reduction by SI-All with respect to SI-Pau.*

ID	# Train	# Test	SD-Pau	SI-Pau	SI-All	RER	ID	# Train	# Test	SD-Pau	SI-Pau	SI-All	RER
me013	115.2k	51.2k	9.20	8.93	**8.36**	6.29	mn052	10.7k	3.8k	8.96	8.93	**8.29**	7.17
me011	50.6k	24.8l	7.27	7.47	**6.61**	11.50	mn021	9.6k	4.1k	9.97	8.23	**8.01**	2.64
fe008	50.6k	22.6k	9.08	8.92	**8.53**	4.37	me003	9.3k	3.6k	6.43	6.18	**5.83**	5.79
fe016	32.0k	15.4k	10.30	10.15	**9.62**	5.18	mn005	7.7k	3.1k	7.73	8.74	**7.73**	11.57
mn015	31.9k	14.7k	9.23	8.69	**7.99**	8.06	me045	8.1k	2.4k	7.99	7.95	**7.20**	9.37
me018	31.8k	14.7k	8.44	8.30	**7.74**	6.72	me025	7.7k	2.4k	8.78	8.74	**8.32**	4.85
me010	26.1k	12.6k	9.25	9.25	**8.30**	10.21	me006	6.9k	1.5k	11.18	10.72	**9.86**	7.98
mn007	27.2k	10.1k	10.84	11.53	**10.71**	7.10	**me026**	5.2k	2.5k	9.47	**7.94**	**7.94**	**0.00**
mn017	21.0k	7.1k	8.55	8.67	**8.03**	7.35	me012	5.3k	2.1k	9.18	8.85	**8.66**	2.11
mn082	13.3k	4.2k	10.17	9.76	**9.00**	7.82	fn002	5.9k	1.5k	10.19	11.26	**9.79**	**13.09**

Table 2: *Comparison of performance of SI, SD, and SI+SD models [E %]. DP stands for the percentage of the training data available for training the SD model (relative to the SI model), CER is the chance error rate, λs are interpolation weights (corresponding to SI) estimated on jackknifed test data, and RER is the relative error rate reduction with respect to CER; IDs of speakers whose SD or SI+SD outperformed the SI model are shown in boldface, as is the best result for each speaker, * indicates that the improvement is significant by a Sign test.*

ID	DP	CER	SI	SD	SI+SD	λs	RER	ID	DP	CER	SI	SD	SI+SD	λs	RER
me013	20.3	13.66	**8.36**	8.47	8.39	1.0,0.9	38.8	mn052	1.9	16.53	**8.29**	8.64	8.32	0.8,0.8	49.8
me011*	8.9	16.09	6.61	6.60	**6.41**	0.5,0.5	60.2	mn021	1.7	13.06	**8.01**	9.27	8.08	0.9,0.8	38.6
fe008	8.9	13.79	**8.53**	8.55	8.55	0.7,0.9	38.1	me003	1.6	13.36	**5.83**	6.57	5.83	1.0,1.0	56.4
fe016*	5.6	14.54	9.62	9.55	**9.52**	0.7,0.7	34.5	**mn005***	1.4	12.99	7.73	**7.15**	7.18	0.0,0.1	45.0
mn015	5.6	14.41	7.99	8.47	**7.96**	0.8,0.8	44.8	me045	1.4	22.31	**7.20**	7.62	7.29	0.8,0.7	**67.7**
me018	5.6	17.22	**7.74**	8.09	7.74	1.0,1.0	55.0	me025	1.4	18.07	**8.32**	8.95	8.32	1.0,1.0	54.0
me010	4.6	14.11	8.30	**8.20**	8.30	0.6,0.5	41.9	me006	1.2	19.46	**9.86**	10.65	9.99	0.8,1.0	49.3
mn007*	4.8	20.52	10.71	10.47	**10.19**	0.6,0.4	50.3	me026	0.9	11.28	**7.94**	8.99	7.94	1.0,1.0	29.6
mn017	3.7	15.05	**8.03**	8.06	8.03	1.0,1.0	46.7	me012	0.9	16.21	**8.66**	8.76	8.66	1.0,1.0	46.6
mn082	2.3	11.17	**9.00**	9.62	9.02	0.8,1.0	**19.4**	**fn002**	1.0	19.71	9.79	10.52	**9.32**	0.7,0.7	52.7

the SI model. That many speakers do not benefit from SD modeling is consistent with their being well described by the SI model. That is, there are most likely some consistent ways that people behave prosodically, but for some speakers who deviate from these norms, speaker-dependent modeling can be of value.

The interpolation weights λ differ across speakers. As expected, the improved speakers have on average relatively lower weights for SI model. In contrast, there are some speakers who have $\lambda = 1$, and thus do not use the SD information for interpolation at all. Note that it is possible for the SI+SD to perform worse than the SI model, because weights are estimated on fairly small amounts of data that are separate from the data on which the model is tested. It is also interesting that the reduction with respect to chance error varies widely across speakers (from 19.4 to 67.7%), but in a manner uncorrelated with training set size.

Figure 1 displays relative *feature usage* statistics for those speakers for whom there is an improvement using the SD model. Feature usage [3] reflects the number of times a feature is queried in a tree, weighted by the number of samples it affects at each node. Total feature usage within a tree sums to 1; results here are based on averaging results over multiple trees. We grouped the prosodic features into five nonoverlapping groups: pause at the boundary in question, duration, pitch, energy, and "near pause" (describing pauses associated with the previous and the following word boundary positions). We compare the SD feature usage distribution with the SI distribution, for native speakers (top graph of the figure) and nonnative speakers (bottom graph).

The two natives show very similar usage to each other and to the SI model. However, as we saw earlier, SD models improve their results significantly. This suggests that even when general feature usage patterns for a talker are similar to those of the SI model, specific features and/or feature thresholds may still be better modeled by training on the specific speaker. Given only two native speakers showing improvements here; it is likely that not all native speakers show the same pattern, but this is a question for further research on a larger data set.

Feature usage for nonnative speakers, on the other hand, looks quite different. Speakers differ from each other, as well as from the SI pattern. Although more research is needed before drawing conclusions, this finding is nevertheless consistent with stylistic differences between nonnative speakers and an overall SI model, in prosodic marking of DA boundaries. Obvious next question would be whether improvement depends on native language, proficiency in English, or degree of perceived accent. Our sample of nonnative speakers is too small to examine these questions, however, we do note that of three native German speakers, all highly proficient in English, one speaker improved from individual modeling while two others did not. Of three Spanish speakers, all moderately pro-

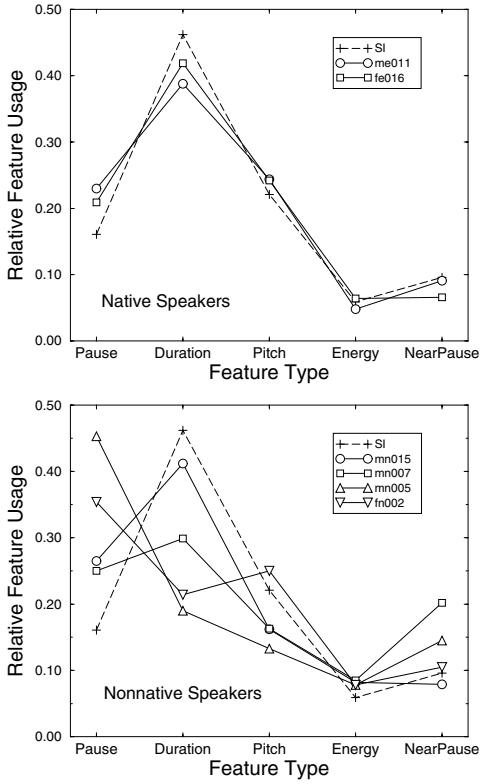

Figure 1: *Relative usage of groups of prosodic features for native (top) and nonnative (bottom) speakers who improved using SD information*

ficient, two improved and one did not.

Overall, duration is the most used feature group, followed by pause and pitch features. As shown, there is rather limited use of energy features, possibly because of issues in normalization (for channel effects), although channel-based normalization was attempted as described earlier. The "near pause" group, which can reflect both nearby hesitation but also short DAs (such as one-word backchannels), is used quite rarely overall as well.

4. Conclusions

We investigated speaker-specific prosodic modeling for DA segmentation in meetings, and found that overall, prosodic features beyond pause provide significant benefit over the pause-only features used in previous work. We further found that for about 30% of the speakers studied, interpolating the large, speaker-independent prosodic model with a much smaller prosodic model trained only on that talker's speech yielded improvements. Feature analysis, while preliminary given the number of speakers, suggests that nonnative speakers may differ from native speakers in overall feature usage patterns associated with DA boundaries.

An important question for future work is to explore what factors predict whether speaker-dependent modeling will benefit a particular speaker, since it did not benefit all speakers. The absolute amount of data did not appear to be a predictor in our experiments, although data is certainly necessary for robustness. Additional areas for further research include examination of results using both prosodic and language models, development of other adaptation methods, and exploration of the clustering of speakers similar in behavior, for greater model robustness.

5. Acknowledgments

This work was supported by the EU 6th FWP ISR Integrated Project AMI (FP6-506811), the DARPA CALO project (NBCHD-030010), NSF project IIS-0121396, DARPA Contract No. HR0011-06-C-0023, and the Ministry of Education of the Czech Republic (project No. 1M0567). The views expressed are those of the authors, and not the funding agencies.

6. References

[1] Armstrong, S. et al.: "Natural Language Queries on Natural Language Data: A Database of Meeting Dialogues," in *Proc. NLDB*, Burg/Cottbus, Germany, 2003

[2] Waibel, A. et al.: "Advances in Automatic Meeting Record Creation and Access," in *Proc. ICASSP*, Salt Lake City, USA, 2001

[3] Shriberg, E., Stolcke, A., Hakkani-Tur, D., Tur, G.: "Prosody-based Automatic Segmentation of Speech into Sentences and Topics," in *Speech Communication*, vol. 32, no. 1–2, p. 127–154, 2000

[4] Warnke, V., Kompe, R., Niemann, H., Nöth, E.: "Integrated Dialog Act Segmentation and Classification Using Prosodic Features and Language Models" *in Proc.* EUROSPEECH 97, pp. 207–210, Rhodes, Greece, 1997

[5] Christensen, H., Gotoh, Y., Renals, S.: "Punctuation Annotation Using Statistical Prosody Models," in *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, USA, 2001

[6] Huang, J., Zweig, G.: "Maximum Entropy Model for Punctuation Annotation from Speech," in *Proc. ICSLP 2002*, Denver, 2002

[7] Kim, J.H., Woodland, P.: "A Combined Punctuation Generation and Speech Recognition System and Its Performance Enhancement Using Prosody," in *Speech Communication*, vol. 41, no. 4, pp. 563–577, 2003

[8] Liu, Y., Stolcke, A., Shriberg, E., Harper, M.: "Using Conditional Random Fields for Sentence Boundary Detection in Speech," in *Proc. ACL* , Ann Arbor, 2005

[9] Ang, J., Liu, Y., Shriberg, E.: "Automatic Dialog Act Segmentation and Classification in Multiparty Meetings," in *Proc. ICASSP-2005*, Philadelphia, 2005

[10] Ji, G., Bilmes, J.: "Dialog Act Tagging Using Graphical Models," in *Proc. ICASSP-2005*, Philadelphia, 2005

[11] Zimmermann, M., Stolcke, A., Shriberg, E.: "Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings," in *Proc. ICASSP-2006*, Toulouse, France, 2006

[12] Kolář, J., Shriberg, E., Liu, Y.: "Using Prosody for Automatic Sentence Segmentation of Multi-Party Meetings," in *Text, Speech and Dialogue (TSD) 2006*, Brno, Czech Republic, 2006

[13] D. Hirst and A. Di Cristo (eds.): *Intonation Systems*, Cambridge University Press, Cambridge, 1998.

[14] Ostendorf, M. and Veilleux, N.: "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Locations," in *Computational Linguistics*, 20 (1), pp. 27-54, 1994.

[15] Janin, A. et al.: "The ICSI Meeting Corpus," in *Proc. ICASSP-2003*, Hong Kong, 2003

[16] Dhillon, R. et al.: *Meeting Recorder Project: Dialog Act Labeling Guide*, ICSI Technical Report TR-04-02, ICSI, Berkeley, USA, 2004

[17] Buckow, J. et al.: "Fast and Robust Features for Prosodic Classification," in *Proc. TSD'99 Marienbad*, Springer Verl., Berlin, 1999

[18] Breiman, L.: "Bagging Predictors," in *Machine Learning* 24(2), pp. 123–140, 1996

[19] Liu, Y. et al.: "A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech," in *Computer Speech and Language*, To Appear