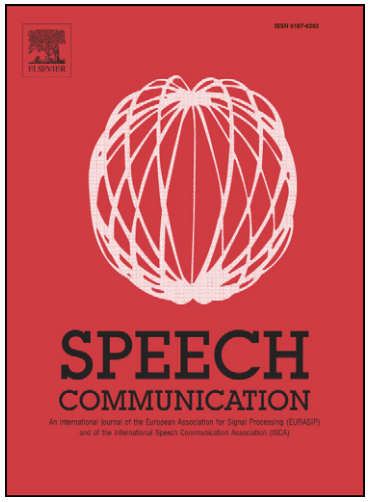
Speaker Adaptation of Language and Prosodic Models for Automatic Dialog Act Segmentation of Speech

Jáchym Kolář, Yang Liu, Elizabeth Shriberg



Speaker Adaptation of Language and Prosodic Models for Automatic Dialog Act Segmentation of Speech

Jáchym Kolá ř, Yang Liu, Elizabeth Shriberg

PII: S0167-6393(09)00163-0
DOI: 10.1016/j.specom.2009.10.005
Reference: SPECOM 1841

To appear in: *Speech Communication*

Received Date: 14 July 2009
Revised Date: 20 October 2009
Accepted Date: 20 October 2009

Please cite this article as: Kolá ř, J., Liu, Y., Shriberg, E., Speaker Adaptation of Language and Prosodic Models for Automatic Dialog Act Segmentation of Speech, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.10.005

Speaker Adaptation of Language and Prosodic Models for Automatic Dialog Act Segmentation of Speech

Jáchym Kolář*[,a], Yang Liu[b], Elizabeth Shriberg[c,d]

[a]*Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic*
[b]*Department of Computer Science, University of Texas at Dallas, Richardson, TX, U.S.A.*
[c]*Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, U.S.A.*
[d]*International Computer Science Institute, Berkeley, CA, U.S.A.*

Abstract

Speaker-dependent modeling has a long history in speech recognition, but has received less attention in speech understanding. This study explores speaker-specific modeling for the task of automatic segmentation of speech into dialog acts (DAs), using a linear combination of speaker-dependent and speaker-independent language and prosodic models. Data come from 20 frequent speakers in the ICSI meeting corpus; adaptation data per speaker ranges from 5k to 115k words. We compare performance for both reference transcripts and automatic speech recognition output. We find that: (1) speaker adaptation in this domain results both in a significant overall improvement and in improvements for many individual speakers, (2) the magnitude of improvement for individual speakers does not depend on the amount of adaptation data, and (3) language and prosodic models differ both in degree of improvement, and in relative benefit for specific DA classes. These results suggest important future directions for speaker-specific modeling in spoken language understanding tasks.

Key words: Spoken language understanding, Dialog act segmentation, Speaker adaptation, Prosody modeling, Language modeling

1. Introduction

The general idea of model adaptation to a particular talker has successfully been used in the cepstral domain for speech recognition, for example by Gauvain and Lee (1994) and Gales (1998). However, less is known about speaker adaptation for spoken language *understanding*. This paper explores the question of speaker adaptation of generic models for a language understanding task. We focus on speaker-specific modeling for one spoken language understanding task, automatic dialog act (DA) segmentation of speech. This task is important since standard automatic speech recognition (ASR) systems output only a raw stream of words, leaving out important structural information such as locations of sentence or DA boundaries. Such locations are overt in standard text via punctuation and capitalization, but are "hidden" in speech. As shown by a number of studies, the absence of sentence or DA boundaries in speech transcripts causes difficulties for both humans and computers.

Effects on human sentence processing were studied by Jones et al. (2003), who demonstrated that sentence breaks are critical for readability of speech transcripts. Moreover, a lack of sentence segmentation can make the meaning of some utterances ambiguous. To take an extreme case, if an automatic speech recognizer outputs the stream of words *"no rooms are available"*, it is not clear what was said – whether it was *"No rooms are available."* or *"No. Rooms are available."* In this example, the two possible interpretations have completely opposite meaning. Such cases are relatively rare, but other forms of ambiguity can be much more frequent.

Lack of linguistic unit boundaries also causes sig-

*Corresponding author. Jáchym Kolář, University of West Bohemia, Department of Cybernetics, Univerzitní 8, 306 14 Plzeň, Czech Republic.
Tel.:+420-377-63-2563, Fax.:+420-377-63-2502.
Email addresses: jachym@kky.zcu.cz (Jáchym Kolář), yangl@hlt.utdallas.edu (Yang Liu), ees@speech.sri.com (Elizabeth Shriberg)

nificant problems for automatic processing. Many natural language processing (NLP) techniques (e.g., parsing, automatic summarization, information extraction, machine translation) are typically trained on well-formatted input, such as text, and fail when dealing with unstructured streams of words. For instance, Furui et al. (2004) reported that speech summarization improved when sentence boundaries were provided. In the area of parsing, Kahn et al. (2004) achieved a significant improvement in parsing performance when using a more accurate sentence boundary detection system. Furthermore, Matusov et al. (2007) showed that the use of automatically-detected sentence boundaries is beneficial for machine translation.

State-of-the-art approaches to DA segmentation typically use both lexical and prosodic information. Most prior work on this task has focused on identifying effective features or on developing advanced models. Such work has almost exclusively trained aggregate models, representing data pooled over speakers.

In this work, we investigate whether the speakers differ enough from each other in the production of DA boundaries to merit speaker-dependent modeling for this task. We perform speaker adaptation for both language and prosodic models, using a speech corpus of multiparty meetings. The meeting domain is chosen for several reasons. First, we are interested in spontaneous speech, since modeling of idiosyncratic lexical patterns for DA segmentation would not be meaningful for corpora of read speech. We also expect that idiosyncratic prosodic patterns are best seen in spontaneous speech. Second, as in any study of adaptation, it is essential to have enough data to adapt the general model to the specific one. Thus in our case the target domain should have speakers with plenty of speech data, and ideally data from different conversations for purposes of generalization.

For these reasons, we use data from a corpus comprising a series of naturally-occurring meetings. This corpus, like many real-world meetings, has recurring participants, presenting the opportunity for adapting models to the individual talkers. Furthermore, in this corpus, as in other meeting applications, the speakers are known beforehand and are recorded on separate channels. This allows us to focus on the question of inherent contributions from speaker-adaptive modeling, rather than confound results with the issue of speaker separation or recognition.

We ask several questions about speaker variation in lexical and prosodic patterns associated with DA boundaries. First, we ask whether speakers differ enough from overall (speaker-independent) models to benefit from model adaptation using a relatively small amount of their speech. Second, we explore whether the effectiveness of adaptation is correlated with the data amount available for adaptation. If this is not found to be the case, then it would suggest that speakers differ inherently in how well they are characterized by generic models. Third, we investigate whether adaptation performance is dependent on different DA types (for example statements versus questions). Finally, we compare speaker adaptation results for language modeling versus prosodic modeling.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 describes our language and prosodic models for DA segmentation, and the speaker adaptation approach. Section 4 and Section 5 present our experimental setup and discuss the experiment results. Section 6 provides a summary and conclusions.

2. Related Work

General (speaker-independent) methods for automatic detection of linguistic unit boundaries in speech have been studied quite extensively in the past decades. Several different approaches utilizing both textual (lexical or syntactic) and acoustic (prosodic) information have been proposed. The proposed techniques include hidden Markov models (HMMs) (Shriberg et al., 2000; Kim and Woodland, 2003), multilayer perceptrons (Warnke et al., 1997; Srivastava and Kubala, 2003), maximum entropy (Huang and Zweig, 2002; Liu et al., 2004), conditional random fields (Liu et al., 2005; Zimmermann, 2009), support vector machines (Akita et al., 2006; Magimai-Doss et al., 2007), and adaptive boosting (Zimmermann et al., 2006; Kolář et al., 2006b). Syntactic information has been used in (Roark et al., 2006; Favre et al., 2008). Domain adaptation for sentence boundary detection has been studied by Cuendet et al. (2006). However, basically no attention has been paid to *speaker adaptation* of lexical or prosodic models.

For related work, we should mention papers focusing on speaker-dependent modeling for general LM adaptation in speech recognition. Besling and Meier (1995) improved an automatic speech dictation system by speaker LM adaptation based

on the LM fill-up method. Akita and Kawahara (2004) showed improved recognition performance using LM speaker adaptation by scaling the n-gram probabilities with the unigram probabilities estimated via probabilistic latent semantic analysis. Tur and Stolcke (2007) demonstrated that unsupervised within-speaker LM adaptation significantly reduced word error rate in meeting speech recognition.

Even less is known about speaker-specific variation in prosodic patterns, beyond basic F_0 normalization used by Shriberg et al. (2000). Studies in speech synthesis and speaker recognition have used prosodic variation successfully, but to our best knowledge, modeling stylistic prosodic variability for sentence or DA boundary detection has been mentioned only anecdotally in the literature (Ostendorf and Veilleux, 1994; Hirst and Cristo, 1998).

We have already presented preliminary results of this work in two conference papers (Kolář et al., 2006a, 2007). Unlike the two earlier papers, this paper contains more results, analysis, and discussion. It also compares the two adaptation methods (language and prosodic) and investigates effects based on DA types.

3. Modeling and Features

For a given word sequence $W = w_1 w_2 ... w_i ... w_n$, the task of DA segmentation is to determine which interword boundaries correspond to a DA boundary. We label each interword boundary as either a within-unit boundary or a boundary between two DAs. For example, in the utterance "yes we should be done by noon", there are two DAs: "yes" and "we should be done by noon". Here we describe language and prosodic models used in this work, as well as our approach to their speaker adaptation.

3.1. Language Models (LM)

We use a hidden event LM (Stolcke et al., 1998) to automatically detect DA boundaries in the unstructured word sequence. This approach based on the hidden Markov model (HMM) framework has been widely used for sentence and DA segmentation and generally achieves performance comparable to other approaches. The hidden event LM describes the joint distribution of words and DA boundaries, $P_{LM}(W, B)$, where B is the DA boundary sequence corresponding to the word sequence W. In training, the DA boundary is explicitly included as a token in the vocabulary, and an n-gram LM is trained based on the word and DA boundary sequences in the training data. In all the experiments reported herein, we used trigram LMs with modified Kneser-Ney smoothing (Kneser and Ney, 1995) in the SRILM toolkit (Stolcke, 2002).

During testing, a forward-backward algorithm is used to compute the posterior probability $P(B_i|W)$ of a boundary B_i at position i given a word sequence W. Then a decision is made to select B_i with the highest posterior probability $P(B_i|W)$ at each individual boundary. This approach minimizes the expected per-boundary classification error rate.

3.2. Prosodic Features and Models

Our prosodic features are designed to reflect breaks in temporal, intonational, or energy contours. Note that we are interested in speaker differences after normalizing for a speaker's inherent pitch range. The features are extracted from the word-level and phone-level time alignment information obtained from an automatic speech recognizer (Shriberg et al., 2000). This approach computes features directly from the signal, without the need for any human labeling of prosodic events. The prosodic features can be grouped into four broad feature classes: *pause*, *pitch*, *duration*, and *energy*. Features are computed in regions local to each candidate boundary to be classified. That is, we compute prosodic features in a window around each of the arrows in a word sequence $w_1 \downarrow w_2 \downarrow ... w_i \downarrow ... \downarrow w_n$, where the word sequence is from either the reference transcripts or the ASR output. Features reflect prosodic information either before, after, or across the word boundary. To capture local prosodic dynamics, we also use features associated with the previous and the following boundaries.

The pause features capture raw durations of pauses between adjacent words. The duration features include phone-normalized durations of vowels, final rhymes, and words; normalization statistics are generated from the entire training set. We did not use raw duration features, since they correlate with lexical features that should be modeled in a language model. Certain frequent DAs (especially backchannels) have small sets of words, so raw durations may capture those words rather than prosody. Our 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. The pitch features are extracted from

Table 1: Data set sizes for individual speakers. ID=Speaker ID, Train=Training set size, Test-Ref=Test size for REF, Test-ASR=Test size for ASR. All data sizes are presented as numbers of words.

No.	ID	Train	Test-Ref	Test-ASR	No.	ID	Train	Test-Ref	Test-ASR
1.	me013	115.2k	51.2k	43.4k	11.	mn052	10.7k	3.8k	3.5k
2.	me011	50.6k	24.8k	22.9k	12.	mn021	9.6k	4.1k	4.1k
3.	fe008	50.6k	22.6k	19.5k	13.	me003	9.3k	3.6k	3.2k
4.	fe016	32.0k	15.4k	13.9k	14.	mn005	7.7k	3.1k	3.0k
5.	mn015	31.9k	14.7k	13.7k	15.	me045	8.1k	2.4k	2.1k
6.	me018	31.8k	14.7k	13.3k	16.	me025	7.7k	2.4k	1.6k
7.	me010	26.1k	12.6k	11.3k	17.	me006	6.9k	1.5k	1.3k
8.	mn007	27.2k	10.1k	8.4k	18.	me026	5.2k	2.5k	2.3k
9.	mn017	21.0k	7.1k	6.0k	19.	me012	5.3k	2.1k	1.9k
10.	mn082	13.3k	4.2k	3.7k	20.	fn002	5.9k	1.5k	1.4k

F_0 contours stylized by a piece-wise linear function (Sönmez et al., 1998). The energy features are represented by the maximal, minimal, and mean frame-level RMS values, using both raw and per-channel normalized values.

For prosody modeling, we used decision tree classifiers (Breiman et al., 1984) because they have been found in past work to yield good results, and because they offer the advantage of interpretability with respect to individual features. Because DA boundaries are much less frequent than non-DA boundaries, we had to cope with the problem of data skew. To overcome this problem and to decrease classifier variance, we use a combination of ensemble sampling with bagging (Liu et al., 2006). Since the trees were trained on bagged ensembles downsampled to equal class priors, when applying the classifiers on (the imbalanced) test data, the output posteriors, $P(B|AP)$ (where AP is acoustic prosodic features), were adjusted to take into account the original class priors.

We first developed a large set of candidate prosodic features, and then performed feature selection to identify a small set of useful features, in two steps. First, for each of the broad prosodic feature categories, we selected those features with a feature usage statistic (measured by the number of times the feature is used in the decision trees, see *Analysis 1* in Section 5.2 for more details) higher than a predefined threshold. Then using these features, we performed leave-one-out feature selection and removed a feature if its deletion did not yield any performance loss. The selection was performed in speaker-independent fashion. We also tried a speaker-dependent selection, but it did not yield overall performance improvement. The final prosodic feature set contained 32 features. Among them, 16 correspond to *duration*, 10 to *pitch*, 3 to *energy*, and 3 to *pause*. Importance of individual feature groups in terms of feature usage will be discussed in *Analysis 1* in Section 5.2.

3.3. Speaker Adaptation Approach

To adapt the generic speaker-independent models to a particular speaker, we employ an interpolation approach. The same approach is used for both language and prosodic model adaptation. The speaker-adapted (SA) result is obtained from a linear combination of posterior probabilities from the speaker-independent model SI, which is trained from all available data, and a speaker-dependent model SD, which is trained using only data from the target talker, that is

$$P_{SA}(B|X;\lambda) = \lambda P_{SI}(B|X) + (1-\lambda) P_{SD}(B|X) \quad (1)$$

where X denotes either the observed word context or prosodic feature vector, depending on the adapted model type, $P_{SI}(B|X)$ and $P_{SD}(B|X)$ denote speaker-independent and speaker-dependent posteriors, and λ is a weighting factor that is empirically optimized on development data. Note that the SD data is already contained in the SI data for training. Therefore, for the case of LM adaptation, the interpolation does not help reduce out-of-vocabulary rate. It rather gives a larger weight to n-grams observed in the data corresponding to a particular speaker and is expected to better suit this speaker. Analogous assumptions also motivate prosodic model adaptation.

4. Data and Experimental Setup

4.1. Speech Database and Dialog Act Annotation

All experiments presented in this paper are evaluated using the ICSI meeting corpus (Janin et al., 2003). The database contains approximately 72 hours of multichannel conversational speech data and associated human transcripts, manually annotated for DAs (Dhillon et al., 2004). The DA annotation distinguishes five broad DA classes:

- *Backchannels* (B) – Responses to a speaker who has the floor as that speaker is talking; such responses generally do not elicit feedback
  Spk 1: "*We'll start with the presentation –*"
  Spk 2: "*Uh-huh.*"
  Spk 1: "*– and then have lunch.*"
- *Disruptions* (D) – Utterances that are indecipherable, abandoned, or interrupted
  "*Tell me about the –* | *Do you hear me?*"
- *Floor grabbers/holders* (F) – DAs pertaining to mechanisms of grabbing or maintaining the floor
  Spk 1: "*I am sure about it.* | *I can –*"
  Spk 2: "*Well I –* | *It's not so easy.*"
- *Questions* (Q) – Interrogative DAs
  "*Are you happy?*"
- *Statements* (S) – Standard declarative DAs
  "*She will arrive tomorrow.*"

A small number of DAs were not assigned into any of the above presented classes. These were mostly pre- or post-meeting chatter or utterances that were mumbled or could not be understood for some reason.

4.2. Reference and ASR Transcripts

Two different test conditions are used for evaluation: reference transcripts (REF) and automatic speech recognition output (ASR). Recognition results were obtained using the state-of-the-art SRI ASR system, originally developed for the conversational telephone speech domain (Stolcke et al., 2006). The recognizer was trained using no acoustic data or transcripts from the analyzed meeting corpus. To represent a fully automatic system, we also used automatic speech/nonspeech segmentation. Word error rates for this difficult data are still quite high; the ASR system performed at 38.2% (on the whole corpus). To generate the "reference" DA boundaries for the ASR words, we aligned the reference setup to the recognition output with the constraint that two aligned words could not occur further apart than a fixed time threshold (0.5 sec).

4.3. Experimental Setup

For our experiments, we selected the top 20 speakers in terms of each speaker's total number of words. Each speaker's data was split into a training set (∼70% of data) and a test set (∼30%), with the caveat that a speaker's recording in any particular meeting appeared in only one of the two sets. Because of data sparsity, especially for the less frequent speakers, we did not use a separate development set, but rather jackknifed the test set in our experiments. In this approach, one half of speaker's test data is used to tune weights for the other half, and vice versa.

The total training set for speaker-independent models (comprising the training portions of the 20 analyzed speakers, as well as all data from 32 other less-frequent speakers) contained 567k words. The total test set contained 204k words for reference test conditions and 180k for ASR test conditions. Data set sizes for individual speakers are shown in Table 1. Speaker identity is described using the official corpus speaker IDs. The first letter of the ID denotes the gender of the speaker ("f" or "m"); the second letter indicates whether the speaker is a native ("e") or nonnative ("n") speaker of English. The speakers displayed in the table are sorted according to the total numbers of words they have in the corpus. The size of training sets available for training of the speaker-dependent models ranges from 5.2k to 115.2k words. Note that the test set sizes for REF and ASR conditions differ since usually there are fewer words in ASR outputs than in the corresponding reference.

4.4. Evaluation Metric

For performance evaluation, we use classification error rate, called Boundary Error Rate (BER) since it refers to the number of interword boundaries in the test set (Shriberg et al., 2000). It is defined as

$$BER = \frac{Ins + Del}{N_W} \quad [\%] \tag{2}$$

where Ins denotes the number of false DA boundary insertions, Del the number of misses, and N_W the number of words in the test set. We also present chance error for every experiment – the error rate

Table 2: DA segmentation error rates for speaker-independent (SI) and speaker-adapted (SA) *language* models for individual speakers in REF conditions [BER %]. Better result for each speaker is shown in boldface, * indicates that the improvement of SA over SI is significant by the Sign test at $p < 0.05$.

ID	Chance	SI	SA	ID	Chance	SI	SA
me013*	13.66%	6.75%	**6.52%**	mn052	16.53%	7.33%	**7.28%**
me011*	16.09%	7.40%	**7.25%**	mn021*	13.06%	6.68%	**5.65%**
fe008*	13.79%	7.51%	**7.16%**	me003	13.36%	8.78%	**8.56%**
fe016*	14.54%	7.35%	**7.18%**	mn005*	12.99%	7.83%	**6.92%**
mn015*	14.41%	8.05%	**7.80%**	me045	22.31%	**8.90%**	**8.90%**
me018*	17.22%	6.64%	**6.45%**	me025	18.07%	8.06%	**7.85%**
me010*	14.11%	7.24%	**6.84%**	me006	19.46%	9.53%	**9.47%**
mn007*	20.52%	7.59%	**7.31%**	me026	11.28%	**5.80%**	**5.80%**
mn017*	15.05%	7.02%	**6.44%**	me012*	16.21%	6.85%	**6.29%**
mn082	11.17%	6.33%	**6.21%**	fn002	19.71%	**10.92%**	11.33%

Table 3: DA segmentation error rates of speaker-independent and speaker-adapted (SA) *language* models for individual speakers in ASR conditions [BER %]. Better result for each speaker is shown in boldface, * indicates that the improvement of SA over SI is significant by the Sign test at $p < 0.05$.

ID	Chance	SI	SA	ID	Chance	SI	SA
me013*	12.16%	8.29%	**8.18%**	mn052*	14.03%	10.87%	**10.17%**
me011*	14.77%	8.81%	**8.51%**	mn021*	10.70%	8.20%	**7.73%**
fe008*	13.32%	9.19%	**8.89%**	me003	12.82%	9.36%	**9.33%**
fe016	13.72%	8.42%	**8.31%**	mn005*	6.81%	11.47%	**10.42%**
mn015*	12.69%	10.16%	**9.84%**	me045	16.21%	**11.08%**	11.27%
me018*	15.09%	8.12%	**7.90%**	me025	16.89%	**14.36%**	**14.36%**
me010*	13.28%	8.39%	**7.91%**	me006*	16.06%	10.94%	**10.40%**
mn007*	14.80%	11.28%	**10.78%**	me026	10.09%	7.35%	**6.88%**
mn017*	11.80%	8.92%	**7.84%**	me012	13.39%	8.68%	**8.31%**
mn082	12.31%	10.37%	**10.10%**	fn002	13.83%	13.40%	**12.89%**

achieved by classifying every word boundary into the class with the highest prior probability (which is "non-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.

5. Results and Discussion

5.1. Results for Language Model Adaptation

Table 2 shows a comparison of DA segmentation performance for the baseline speaker-independent LM and speaker-adapted LMs for individual speakers, using reference transcripts. The results indicate that performance improved for 17 of 20 speakers. Note that it is possible for the SA 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. The degree of the improvement varies across particular speakers. For 12 talkers, the improvement was statistically significant at $p < 0.05$ using the Sign test.

Table 3 reports the corresponding results for the ASR conditions. The results show that 18 speakers improved by LM speaker adaptation. Of the two speakers that did not improve, one also did not improve for the REF condition. The improvement was statistically significant at $p < 0.05$ for 12 of the 18 improved speakers.

An important finding from these results is that for both test conditions, the relative error reduction achieved by speaker adaptation is not correlated with the amount of adaptation data. This finding suggests that speakers differ inherently in how similar they are to the generic speaker-independent LM. Some talkers probably differ more and thus show more gain, even with less data available for model adaptation.

An overall comparison of performance of baseline

Table 4: Overall DA segmentation error rates of speaker-independent (SI) and speaker-adapted (SA) *language* models in REF and ASR conditions [BER %]

Model	REF	ASR
Chance	15.02%	13.19%
SI	7.30%	9.06%
SA	**6.99%**	**8.76%**

speaker-independent and speaker-adapted LMs is presented in Table 4. In total, the test set contains 204k words for REF and 180k words for ASR conditions. The results show that for both conditions, speaker-adapted LMs outperform the baseline. The overall improvements by LM speaker adaptation for both conditions are statistically significant at $p < 10^{-15}$, using the Sign test. Also note that the average value of the interpolation weight λ (in Equation 1) across all speakers is 0.49, so the SI and SD components are approximately balanced in the speaker-adapted LMs.

To better understand the effects of speaker adaptation on model performance, we also analyzed the results broken down by DA types as described in Section 4.1. The results for individual DA types are shown for both test conditions in Table 5. Along with SI and SA error rates, the table shows numbers of words in each DA group and relative error changes by speaker adaptation. Negative relative numbers indicate error rate reduction, and positive numbers mean increased errors.

In REF conditions, LM speaker adaptation benefited all DA types. The largest relative improvement was achieved for B. Improvements were much smaller for the other types (S, Q, F, and D). From these results, we can infer that B is the class that has the most speaker-idiosyncratic lexical patterns associated with DA boundaries. The results for ASR show some differences in comparison with REF, especially for B. Unlike REF, speaker adaptation resulted in an error increase for B in ASR conditions. The discrepancy may be explained by the largely different numbers of backchannels in REF and ASR transcripts; backchannels are often missed by the ASR system. We notice that the number of backchannels in ASR conditions is 6 times lower than that in REF. We hypothesize that those backchannels that were recognized correctly are also the ones that are less likely to be speaker specific since the frequent backchannels have higher probabilities in the LM used for speech recognition. The highest error reduction by speaker adaptation in ASR conditions was achieved for S, followed by F, Q, and D. The improvements for these types are slightly less than those in the REF condition, possibly due to ASR errors.

5.2. Results for Prosodic Model Adaptation

Table 6 compares performance of SI and SA prosodic models for individual speakers in REF conditions. The results indicate that the SA model is better than the SI model for 7 of the 20 speakers. For 6 other speakers, SI and SA achieved identical results because λ was estimated as 0. Using the Sign test, 4 of the 7 improved speakers showed improvements significant at $p < 0.05$ or better; one speaker (fn002) was marginally significant at $p < 0.10$.

Although only some speakers show these improvements (while some others show rather poor results from SA 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 the SI model. That a number of speakers do not benefit from SA 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 more from these norms, speaker-dependent modeling is an important direction to investigate for further improvements.

Table 7 presents results for ASR conditions. The results indicate that 15 speakers improved by prosodic model adaptation. Although this is more speakers than improved for the REF condition, the magnitudes of improvement were relatively smaller for the ASR than for the REF condition. Note that 6 of the 7 speakers who improved in REF also improved in ASR conditions. Also note that the speakers who improved in ASR but not in REF have on average relatively smaller improvements than those who improved in both conditions. Similar to our findings for LM adaptation, the error reduction of the SA model with respect to SI varies across speakers in a manner uncorrelated with amount of adaptation data. Using the Sign test, only one speaker (mn007) showed a gain significant at $p < 0.05$ in ASR. One other speaker (mn005) showed a marginally significant gain ($p < 0.10$). The variance between REF and ASR may be explained, at least in part, by the differences in the test sets caused by ASR errors. For

Table 5: DA segmentation error rates for SI and SA *language* models by DA type. B – Backchannel, D – Disruption, F – Floor grabber/holder, Q – Question, S – Statement; N_W – Number of test words for the DA type, REL – Relative error change after adaptation (negative numbers indicate error reduction, positive numbers error increase).

	REF				ASR			
DA	**N_W**	**SI**	**SA**	**REL**	**N_W**	**SI**	**SA**	**REL**
B	3.6k	9.19%	7.70%	-16.16%	0.6k	18.40%	18.91%	+2.78%
D	20.9k	12.50%	12.33%	-1.38%	17.6k	14.11%	13.98%	-0.89%
F	4.3k	38.00%	36.81%	-3.13%	3.0k	43.45%	42.61%	-1.93%
Q	15.4k	6.30%	6.02%	-4.43%	13.7k	8.82%	8.69%	-1.49%
S	157.3k	5.75%	5.47%	-4.82%	143.3k	7.65%	7.31%	-4.40%

example, the ASR data contain a smaller number of backchannels, which are more difficult for speaker-specific prosodic modeling (as shown below in *Analysis 2*). However, the results indicate that there is more variance in speaker-dependent prosodic modeling than in speaker-dependent language modeling. Furthermore, we compare the average values of λ between speaker-adapted language and prosodic models. While the average λ across all speakers for LM adaptation is 0.49, it is 0.67 for prosodic model adaptation. Thus, the SD component (1-λ in Eq 1) is less represented in the SA prosodic models than in the SA language models.

The overall results summed over all 20 speakers are shown for both test conditions in Table 8. In REF tests, the overall best performance is from the SA model. The improvement of SA over SI in reference conditions is statistically significant at $p < 0.001$ using the Sign test. Given the absolute difference in BER, the relatively high level of significance may seem a little surprising, but there exist two explanatory reasons. First, the overall test set is quite large (204k words). Second, if the SD information does not benefit a particular speaker, it often gets zero weight in the interpolation model. Thus, both SA and SI show identical output for many unimproved speakers, which decreases variance between the overall outputs and consequently supports statistical significance in the Sign test. In ASR conditions, the SA model was also the best performing one. The superiority of SA over SI was significant at $p < 0.05$. The relative gain by prosodic model speaker adaptation in ASR conditions was higher than that achieved in REF conditions.

Analysis 1: Prosodic feature usage

To better understand the results, we analyzed relative prosodic feature usage for the speakers who improved in REF conditions. REF was chosen rather than ASR to focus on the aspect of the speaker-dependent prosodic characteristics rather than analyzing effects of speech recognition errors. The metric "feature usage" (Shriberg et al., 2000) 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. The statistics here are based on averaging results over multiple trees generated in bagging. For this analysis, the prosodic features were grouped into five nonoverlapping groups: pause at the boundary in question, duration, pitch, energy, and "near pause" (pause durations at the preceding and the following interword boundaries). We compare the SD feature usage distribution with the SI distributions. In addition, we divide the speakers into two categories: native English speakers and non-native speakers. The relative feature usage statistics from this analysis are shown in Figure 1.

The three native speakers show very similar usage to each other and to the SI model. However, as we saw earlier, SA improves their results. 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 that this analysis is based on only three native speakers (their SA model improved over SI), it is possible that not all native speakers show the same pattern. 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 final conclusions, this finding is nevertheless consistent with stylistic differences between nonnative speakers and an overall SI model, in prosodic marking of DA boundaries. An obvious next question would be whether improvement depends on native

Table 6: DA segmentation error rates for SI and SA *prosodic* models in REF conditions [BER %]. Better result for each speaker is shown in boldface, * indicates that the improvement of SA over SI is significant at $p < 0.05$ using the Sign test.

ID	Chance	SI	SA	ID	Chance	SI	SA
me013	13.66%	**8.36%**	8.39%	mn052	16.53%	**8.29%**	8.32%
me011*	16.09%	6.61%	**6.41%**	mn021	13.06%	**8.01%**	8.08%
fe008	13.79%	**8.53%**	8.55%	me003	13.36%	**5.83%**	**5.83%**
fe016*	14.54%	9.62%	**9.52%**	mn005*	12.99%	7.73%	**7.18%**
mn015	14.41%	7.99%	**7.96%**	me045	22.31%	**7.20%**	7.29%
me018	17.22%	**7.74%**	**7.74%**	me025	18.07%	**8.32%**	**8.32%**
me010	14.11%	8.30%	**8.20%**	me006	19.46%	**9.86%**	9.99%
mn007*	20.52%	10.71%	**10.19%**	me026	11.28%	**7.94%**	**7.94%**
mn017	15.05%	**8.03%**	**8.03%**	me012	16.21%	**8.66%**	**8.66%**
mn082	11.17%	**9.00%**	9.02%	fn002	19.71%	9.79%	**9.32%**

Table 7: DA segmentation error rates for SI and SA *prosodic* models in ASR conditions [BER %]. Better result for each speaker is shown in boldface, * indicates that the improvement of SA over SI is significant at $p < 0.05$ using the Sign test.

ID	Chance	SI	SA	ID	Chance	SI	SA
me013	12.16%	8.41%	**8.37%**	mn052	14.03%	8.96%	**8.70%**
me011	14.77%	7.01%	**6.77%**	mn021	10.70%	**7.56%**	**7.56%**
fe008	13.32%	9.04%	**8.90%**	me003	12.82%	6.40%	**6.27%**
fe016	13.72%	**9.40%**	9.41%	mn005	6.81%	6.78%	**5.90%**
mn015	12.69%	8.09%	**7.96%**	me045	16.21%	7.93%	**7.89%**
me018	15.09%	**8.16%**	8.18%	me025	16.89%	11.14%	**10.83%**
me010	13.28%	8.62%	**8.30%**	me006	16.06%	10.94%	**10.63%**
mn007*	14.80%	10.74%	**10.01%**	me026	10.09%	7.91%	**7.61%**
mn017	11.80%	**8.04%**	**8.04%**	me012	13.39%	**8.99%**	**8.99%**
mn082	12.31%	8.03%	**8.00%**	fn002	13.83%	9.74%	**9.46%**

Table 8: Overall DA segmentation error rates for speaker-independent (SI) and speaker-adapted (SA) *prosodic* models in REF and ASR conditions [BER %]

Model	REF	ASR
Chance	15.02%	13.19%
SI	8.25%	8.41%
SA	**8.19%**	**8.26%**

language, proficiency in English, or degree of perceived accent. The sample of nonnative speakers is too small to examine these questions. In our preliminary analysis, we found that among the three native German speakers in our data set, all highly proficient in English, one speaker improved from individual modeling while two others did not. Of the three Spanish speakers, all moderately proficient, two improved and one did not. Further studies are still needed to examine other factors.

Analysis 2: DA types, prosody and LM

We performed analysis of prosodic adaptation benefits for individual DA types. The results are displayed in Table 9, for both REF and ASR conditions. In REF conditions, the highest improvement from speaker adaptation was achieved for D, followed by Q and S. On the other hand, speaker adaptation resulted in worse results for F, and especially for B. The results for ASR conditions are quite similar, with slightly better improvement for most DA types. Speaker adaptation benefited D, S, and Q, but hurt F and B. The effects of prosodic adaptation on B and F are almost opposite to those of LM adaptation. For example, in REF conditions, B had the highest improvement by LM adaptation, but the highest performance drop by prosodic adaptation. Similarly, F shows improvement for LM but not for prosody. Thus, we conclude that, overall, prosodic model speaker adaptation has different improvement patterns than LM speaker adaptation.

Another obvious question is whether the speakers who benefit from prosodic model adaptation

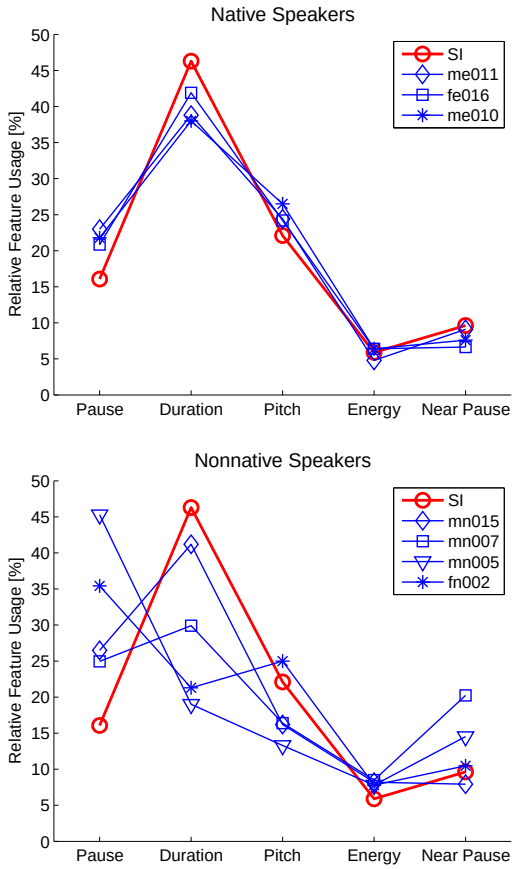


Figure 1: Relative usage of prosodic feature groups for native (top) and nonnative (bottom) speakers who improved from SA in REF conditions

also benefit from LM adaptation. Of the 7 speakers who improved by prosodic adaptation in REF conditions, 6 also improved by LM adaptation. The number of improved speakers in common may seem high, but it approximately corresponds to the chance agreement based on the counts of improved speakers in both sets (5.95 speakers). Similarly, of the 15 speakers who improved by prosodic adaptation in ASR conditions, 13 improved also by LM adaptation, while the chance agreement is 13.5 speakers. These numbers indicate that there was no apparent correlation between speakers' idiosyncrasy in prosodic and lexical patterns associated with DA boundaries.

6. Summary and Conclusions

We have investigated speaker-specific language and prosodic modeling for automatic DA segmentation in multiparty meetings. The method was evaluated on 20 frequent speakers with amounts of data available for speaker-dependent modeling ranging from 5k to 115k words. In the first set of experiments, we explored speaker adaptation of hidden event language models. Improvements were found for 17 of the 20 speakers using reference transcripts, and for 18 of the 20 speakers using automatic transcripts. It was not expected that all speakers may improve because some speakers may be close to the SI model. We also measured overall results, summed over all 20 speakers. These results showed that we achieved a statistically significant improvement over the baseline LM for both test conditions.

In the second set of experiments for speaker adaptation of prosody model, we observed improved results for 7 of the 20 speakers in reference conditions, and for 15 of the 20 speakers in ASR conditions. In the ASR conditions, we observed a higher number of improved speakers, but the improvements were relatively smaller than those in the REF conditions. Also note that the speakers who improved in ASR but not in REF have on average relatively smaller improvements than those who improved in both conditions. The variance between REF and ASR may be explained in part by the differences in the test sets caused by ASR errors. However, the results indicate that there is more variance in speaker-dependent prosodic modeling than in language modeling. Overall results, summed over all 20 speakers, indicate modest yet statistically significant improvement by prosodic model adaptation for both test conditions. The relative overall improvement was smaller than that achieved by LM adaptation. Prosodic 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.

For both types of speaker adaptation, improvements were achieved even for some talkers who had only a relatively small amount of data available for adaptation. In addition, the relative error reduction achieved by speaker adaptation was not correlated with the amount of adaptation data. This finding suggests that speakers differ inherently in how similar they are to the generic models. Some talkers probably differ more and thus show more gain,

Table 9: DA segmentation error rates for SI and SA *prosodic* models by DA type. B – Backchannel, D – Disruption, F – Floor grabber/holder, Q – Question, S – Statement; N_W – Number of test words for the DA type, REL – Relative error change after adaptation (negative numbers indicate error reduction, positive numbers error increase).

	REF				ASR			
DA	**N_W**	**SI**	**SA**	**REL**	**N_W**	**SI**	**SA**	**REL**
B	3.6k	6.69%	7.20%	+7.53%	0.6k	19.93%	21.12%	+5.98%
D	20.9k	9.96%	9.78%	-1.87%	17.6k	9.72%	9.33%	-3.97%
F	4.3k	37.81%	38.40%	+1.54%	3.0k	44.92%	45.42%	+1.12%
Q	15.4k	6.32%	6.23%	-1.34%	13.7k	6.38%	6.27%	-1.83%
S	157.3k	7.43%	7.36%	-0.88%	143.3k	7.63%	7.49%	-1.87%

even with less data. We also found that the agreement between speakers who improved by LM adaptation and those who improved by prosodic adaptation is not higher than chance. Thus, there is no apparent correlation between speakers' idiosyncrasy in prosodic and lexical patterns associated with DA boundaries. Further analysis revealed that LM and prosodic speaker adaptation differ in the DA types that they benefit. For example, while LM adaptation benefits backchannels and floor grabbers/holders, prosodic adaptation hurts them.

We conclude that speaker adaptation aids both prosody- and LM-based DA segmentation, and that future work should investigate the potential of speaker-specific modeling for other tasks in spoken language understanding. The techniques for model adaptation to speaker behavior may be extremely helpful for various speech-based applications. Important areas for future extensions of our work include the integration of lexical with prosodic or even multimodal information in a single model, investigating clustering of speakers similar in behavior for greater model robustness, and exploration of unsupervised adaptation approaches.

Acknowledgments

This work was supported by the Ministry of Education of the Czech Republic under projects 1M0567 and 2C06020 at UWB Pilsen, and the NSF grants IIS-0544682 at SRI International and IIS-0845484 at UT Dallas. The views are those of the authors and do not reflect the views of the funding agencies.

References

Akita, Y., Kawahara, T., 2004. Language model adaptation based on PLSA of topics and speakers. In: Proc. INTERSPEECH 2004-ICSLP. Jeju, Korea.

Akita, Y., Saikou, M., Nanjo, H., Kawahara, T., 2006. Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines. In: Proc. INTERSPEECH 2006 - ICSLP. Pittsburgh, PA, USA.

Besling, S., Meier, H.-G., 1995. Language model speaker adaptation. In: Proc. EUROSPEECH. Madrid, Spain.

Breiman, L., Friedman, J., Ohlsen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks Inc., Pacific Grove, CA, USA.

Cuendet, S., Hakkani-Tur, D., Tur, G., 2006. Model adaptation for sentence segmentation from speech. In: Proc. IEEE Workshop on Spoken Language Technology. Aruba.

Dhillon, R., Bhagat, S., Carvey, H., Shriberg, E., 2004. Meeting recorder project: Dialog act labeling guide. Tech. Rep. TR-04-002, ICSI, Berkeley, CA, USA.

Favre, B., Hakkani-Tur, D., Petrov, S., Klein, D., 2008. Efficient sentence segmentation using syntactic features. In: Proc. IEEE Workshop on Spoken Languge Technology. Goa, India.

Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C., 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech and Audio Processing 12 (4), 401–408.

Gales, M., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language 12, 75–98.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture obsevation of Markov chains. IEEE Transactions on Speech and Audio Processing 2 (2), 291–298.

Hirst, A., Cristo, A. D. (Eds.), 1998. Intonation Systems. Cambridge University Press.

Huang, J., Zweig, G., 2002. Maximum entropy model for punctuation annotation from speech. In: Proc. of ICSLP 2002. Denver, CO, USA.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C., 2003. The ICSI meeting corpus. In: Proc. ICASSP. Hong Kong.

Jones, D., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D., Zissman, M., 2003. Measuring the readability of automatic speech-to-text transcripts. In: Proc. EUROSPEECH. Geneva, Switzerland.

Kahn, J. G., Ostendorf, M., Chelba, C., 2004. Parsing conversational speech using enhanced segmentation. In: Proc. HLT-NAACL. Boston, MA, USA.

Kim, J. H., Woodland, P., 2003. A combined punctuation generation and speech recognition system and its performance enhancement using prosody. Speech Communica-

tion 41 (4), 563–577.

Kneser, R., Ney, H., 1995. Improved backing-off for M-gram language modeling. In: Proc. ICASSP. Detroit, MI, USA.

Kolář, J., Liu, Y., Shriberg, E., 2007. Speaker adaptation of language models for automatic dialog act segmentation of meetings. In: Proc. INTERSPEECH 2007. Antwerp, Belgium.

Kolář, J., Shriberg, E., Liu, Y., 2006a. On speaker-specific prosodic models for automatic dialog act segmentation of multi-party meetings. In: Proc. INTERSPEECH 2006 – ICSLP. Pittsburgh, PA, USA.

Kolář, J., Shriberg, E., Liu, Y., 2006b. Using prosody for automatic sentence segmentation of multi-party meetings. In: Text, Speech and Dialogue. Vol. 4188 of Lecture Notes in Artificial Intelligence. pp. 629–636.

Liu, Y., Chawla, N., Harper, M., Shriberg, E., Stolcke, A., 2006. A study in machine learning from imbalanced data for sentence boundary detection in speech. Computer Speech and Language 20, 468–494.

Liu, Y., Stolcke, A., Shriberg, E., Harper, M., 2004. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In: Proc. EMNLP. Barcelona, Spain.

Liu, Y., Stolcke, A., Shriberg, E., Harper, M., 2005. Using conditional random fields for sentence boundary detection in speech. In: Proc. ACL. Ann Arbor, MI, USA.

Magimai-Doss, M., Hakkani-Tur, D., Cetin, O., Shriberg, E., Fung, J., Mirghafori, N., 2007. Entropy based classifier combination for sentence segmentation. In: Proc. ICASSP. Honolulu, HI.

Matusov, E., Hillard, D., Magimai-Doss, M., Hakkani-Tür, D., Ostendorf, M., Ney, H., 2007. Improving speech translation with automatic boundary prediction. In: Proc. INTERSPEECH 2007. Antwerp, Belgium.

Ostendorf, M., Veilleux, N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary locations. Computational Linguistics 20 (1), 27–54.

Roark, B., Liu, Y., Harper, M., Stewart, R., Lease, M., Snover, M., Shafran, I., Dorr, B., Hale, J., Krasnyanskaya, A., Yung, L., 2006. Reranking for sentence boundary detection in conversational speech. In: Proc. ICASSP. Toulouse, France.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G., 2000. Prosody-based automatic segmentation of speech into sentences and topics. Speech Communication 32 (1-2), 127–154.

Sönmez, K., Shriberg, E., Heck, L., Weintraub, M., 1998. Modeling dynamic prosodic variation for speaker verification. In: Proc. ICSLP. Sydney, Australia.

Srivastava, A., Kubala, F., 2003. Sentence boundary detection in Arabic speech. In: Proc. EUROSPEECH. Geneva, Switzerland.

Stolcke, A., Chen, B., Franco, H., Gadde, V. R. R., Graciearena, M., Hwang, M.-Y., Kirchhoff, K., Mandal, A., Morgan, N., Lei, X., Ng, T., Ostendorf, M., Sönmez, K., Venkataraman, A., Vergyri, D., Wang, W., Zheng, J., Zhu, Q., 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. IEEE Transactions on Audio, Speech, and Language Processing 14 (5), 1729 – 1744.

Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tür, G., Lu, Y., 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In: Proc. ICSLP. Sydney, Australia.

Stolcke, A., 2002. SRILM – An extensible language modeling toolkit. In: Proc. ICSLP. Denver, CO, USA.

Tur, G., Stolcke, A., 2007. Unsupervised language model adaptation for meeting recognition. In: Proc. ICASSP. Honolulu, HI, USA.

Warnke, V., Kompe, R., Niemann, H., Nöth, E., 1997. Integrated dialog act segmentation and classification using prosodic features and language models. In: Proc. EUROSPEECH. Rhodes, Greece.

Zimmermann, M., Hakkani-Tur, D., Fung, J., Mirghafori, N., Gottlieb, L., Shriberg, E., Liu, Y., 2006. The ICSI+ multilingual sentence segmentation system. In: Proc. INTERSPEECH 2006 – ICSLP. Pittsburgh, PA, USA.

Zimmermann, M., 2009. Joint segmentation and classification of dialog acts using conditional random fields. In: Proc. INTERSPEECH 2009. Brighton, UK.