Linguistic Patterns in Spontaneous Speech, 213-239
2009-1-025-010-000272-1

Prosodic Similarities of Dialog Act Boundaries Across Speaking Styles*

Elizabeth Shriberg[1,2], Benoit Favre[2], James Fung[2],
Dilek Hakkani-Tür[2], and Sébastien Cuendet[2]
SRI International[1]
International Computer Science Institute[2]

1. Introduction

Spontaneous speech differs in a multitude of ways from read, formal, or laboratory speech (Maclay & Osgood 1959, Goldman-Eisler 1968, Levelt 1983, Biber 1988, Howell & Kadi-Hanifi 1991, Eskenazi 1993, Shriberg 1994, Swerts et al. 1996, Bruce 1995, Hirschberg 1995, Laan 1997). Although the labels "spontaneous" and "read" each reflect an underlying multidimensional space of different genres or "styles" (Eskenazi 1993), generally speaking, spontaneous speech exhibits greater segmental and suprasegmental variability than does read speech (Lieberman et al. 1985, Llisterri & Poch 1991, Wajskop et al. 1992, Kohler 1996, Greenberg 1999, Maekawa 2003, Tseng 2005, Benzeghiba et al. 2007). It is no surprise, then, that automatic speech processing techniques typically have more difficulty with spontaneous than with read speech (Weintraub et al. 1996, Greenberg 1999, McAllister et al. 1998, Riley et al. 1999, Ostendorf 2000, Binnenpoorte et al. 2004, Benzeghiba et al. 2007). For example, an early study in large-vocabulary speech recognition found degraded performance for spontaneous speech even when recording conditions, speaker, and word sequences were held constant (Weintraub et al. 1996). In the study, speakers read transcripts of what they had said in previous spontaneous conversations; the read versions were significantly easier to recognize than the spontaneous originals.

In this paper we focus on comparisons of prosody across speaking styles, specifically for the task of dialog act segmentation. Dialog act segmentation aims to segment the

* This research was supported by the Defense Advanced Research Projects Agency under contracts HI0011-06-C-0023 and NBCHD030010, and by IM2, a research network funded by the Swiss National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. We thank Harry Bratt for feature extraction software and Andreas Stolcke for helpful comments.

continuous speech stream into dialog act units, i.e., to find the boundaries of dialog acts such as statements, questions, and backchannels. The task is of particular importance for speech understanding applications. Dialog act segments are similar to sentence-level segments, which are required for semantic processing techniques, including machine translation, question answering, and information extraction. Most such applications are developed for text input and rely on the assumption that boundaries are marked overtly via punctuation or text formatting (Shriberg & Stolcke 2004, Mrozinski et al. 2006, Makhoul et al. 2005, Hakkani-Tür & Tür 2007).

When the application uses spoken language rather than text, the input is the stream of words produced by an automatic speech recognizer. Recognizer output typically lacks punctuation, and thus the locations of dialog act boundaries need to be recovered automatically. Automatic boundary annotation has been shown in various studies to aid automatic summarization, named entity extraction, machine translation, and part-of-speech tagging (Furui et al. 2004, Matusov et al. 2007, Hillard et al. 2006, Fügen & Kolss 2007, Rao et al. 2007). It has also been shown to aid human readability of the output of automatic speech recognition systems (Jones et al. 2005) and could be used for determining semantically and prosodically coherent boundaries for playback of speech to users in tasks involving audio search.

Studies of automatic dialog act or sentence segmentation have used lexical, syntactic, prosodic, speaker, and time-based features (Warnke et al. 1997, Shriberg et al. 2000, Kim & Woodland 2003, Liu et al. 2005, Ang et al. 2005, Liu et al. 2006, Kolar et al. 2006, Cuendet et al. 2007b, Batista et al. 2007, Dielmann & Renals 2006, Fügen & Kolss 2007, Matusov et al. 2007, Cuendet et al. 2007a). In many such studies, prosodic features have been shown to improve performance over lexical features alone and to have greater robustness than do lexical features to errors in speech recognition output. The complementarity of prosodic to lexical features for this task also lends itself well to procedures such as co-training, which can be useful when only small sets of boundary-labeled data are available (Guz et al. 2007).

To develop prosody features for automatic segmentation, one can look to the rich history of linguistic descriptions of such phenomena. The task of sentence or dialog act boundary detection corresponds most closely to the linguistics literature pertaining to major phrase boundaries. Prosodic properties of major phrase boundaries and related phenomena are discussed by a number of authors (Ladd 1980, Cutler & Ladd 1983, Vaissière 1983, Cruttenden 1986, Couper-Kuhlen 1986, Bolinger 1986, Bolinger 1989, Ladd 1996, Hirschberg 2002). General patterns for English (as well as for many other languages) include a pause at the boundary, preboundary pitch drop, pitch declination over the phrase, postboundary pitch reset, energy contours similar to pitch behavior, preboundary durational lengthening, and voice quality changes. Boundaries are also associated with particular boundary tones. Certain dialog acts are associated with specific patterns; for example, questions are often described as showing a preboundary intonational rise rather than a fall. Further insights, particularly in the area of boundaries at turn-relevant locations, are provided by work in conversation analysis, pragmatics, and discourse analysis (Sacks et al. 1974, Schegloff 1982, Atkinson & Heritage 1984, Couper-Kuhlen & Selting 1996).

Over the past decade, a line of research in computational processing has used a "direct modeling" approach aimed at capturing the phenomena from linguistic descriptions, such as those just mentioned, for use in automatic segmentation and other spoken-language processing tasks (Shriberg 2005). In the direct modeling approach (Shriberg & Stolcke 2004) no human annotation of prosodic events is required. Instead, features are extracted directly from the signal, and a classifier is trained to learn the relationship between the extracted features and the classes to be distinguished for the particular task at hand. Thus, instead of using phonological constructs (e.g., pitch accents or boundary tones) to find dialog act boundaries, the direct modeling approach uses sets of features designed to capture breaks in phrasing, such as pause information, local pitch slopes, or energy differences across candidate boundary locations. The approach generally uses a large set of prosodic features, some highly intercorrelated, and leaves it to a machine classifier to determine how to make best use of the available features. Because prosodic features do not depend on specific words, they offer the possibility of greater generalizability across different speech corpora than do lexical features.

A repeated finding, however, is that when automatic classifiers are used to predict sentence or dialog act boundaries from prosodic features, different sets of features are found to be useful for different corpora and speaking styles. Feature utility is determined by running machine learning experiments using subsets of features, or by determining feature importance when all features are available to the classifier, or via various feature selection approaches. Importantly, selected features vary depending on the data set, even if feature definitions, extraction methods, and classifiers are held constant (Shriberg et al. 2000, Liu et al. 2006, Cuendet et al. 2007a, Cuendet et al. 2007b).[1]

One explanation for the differences in prosodic feature usage across data sets is simply that inherent prosodic cues to dialog act boundaries differ across speaking styles. Given the literature on differences across speaking styles noted earlier, it would not be surprising if speakers adopted different strategies for marking dialog act boundaries prosodically, depending on the speaking context. For automatic processing, this

[1] We focus here on English language data. A separate but related controlled study of prosodic feature selection for three different languages has been recently reported (Fung et al. 2007). That study again finds differences in feature sets across languages, but no claims are made here about language-related prosodic differences.

hypothesis implies that dialog act segmentation approaches should use genre-specific models, thereby requiring some type of matched training data for the genre at hand.

An alternative hypothesis, however, is that there exist inherent prosodic similarities or invariants across styles, but that for reasons having to do with how experiments are conducted, such consistencies have not been immediately obvious. This possibility is of particular interest from an applied perspective. If features that help distinguish boundaries from nonboundaries are qualitatively similar for very different speaking styles, then it should be possible, in principle, to learn robust models that automatically detect dialog act boundaries across genres.

To explore the question of cross-genre prosodic feature similarities in boundary marking, we compare data from two very different speaking contexts—face-to-face meetings and read news broadcasts, as described in more detail in §2.1. The two contexts were chosen because they represent essentially opposite extremes on dimensions of naturalness and level of human-human interaction. The data sets also contain separate groups of speakers.

We examine prosodic features of word transitions for dialog act boundaries and nonboundaries, computed using matched procedures across corpora. We look at the *difference* between the two classes across corpora, as well as the distributions by class across corpora. We break down features by type, since it is conceivable that, for example, duration features may pattern one way but pitch another. In particular, pauses reflect both prosodic and discourse (turn-taking) factors, and the latter is certain to differ between conversational and read speech.

§2 describes the data, dialog acts, features, classifiers, and metrics used. §3 compares results of automatic boundary classification experiments for the two corpora, broken down by feature type. Inherent feature discrimination analyses are then presented in §4, for each of the prosodic feature types (pause, duration, pitch, energy). §§5 and 6 provide a general discussion and conclusions.

2. Method

2.1 Corpora and dialog acts

To represent a spontaneous speaking context at one end of the naturalness dimension, we examine data from the ICSI Meeting Recorder Dialog Act (MRDA) corpus (Shriberg et al. 2004, Janin et al. 2003). The ICSI Meeting corpus is a collection of 75 naturally occurring meetings, including simultaneous multichannel audio recordings and word-level orthographic transcriptions. This corpus has the advantage of being fully hand labeled for dialog acts and their boundaries (Dhillon et al. 2004). Participants knew each other, since they generally met regularly in their working environment. The meetings were not staged scenarios but rather actual meetings with goals related to the everyday research objectives of the participants. Meetings averaged about an hour in length, with a maximum of 103 and minimum of 17 minutes. We use a 73-meeting subset of this corpus that has been used in other studies of the meeting data (Ang et al. 2005, Kolar et al. 2006, Cuendet et al. 2007b), with the same split into 51 training, 11 held-out, and 11 test meetings. The held-out data is used for tuning of model combination and was selected to be roughly matched for distribution of meeting type and general statistics.

For purposes of this study, detailed dialog act annotations (Dhillon et al. 2004) are collapsed into the small set of orthogonal labels shown in Table 1, using a class mapping provided with the MRDA corpus. Each word transition is mapped to either a nonboundary (**n**) or a boundary (**s**, **q**, **b**, or **d**), where abbreviations are defined in the table.[2] Statements comprise the majority of the utterances containing propositional content. Questions include all forms, including yes-no, wh-questions, questions with declarative syntax, and tag questions. Backchannels such as “yeah” and “right” are typically only one or two words long and provide feedback that the listener is attending to what the foreground talker is saying, without taking the floor. Disrupted utterances include both speaker-initiated cut-offs (such as false starts) and cut-offs attributable to interruption from one or more other talkers. In analyses to follow, dialog acts in the meeting data are grouped into two classes, with statements, questions, backchannels, and disrupted utterances forming the “boundary” class. The disruption class shares characteristics with both boundaries and nonboundaries. Looking backward in time, it resembles a nonboundary. But looking forward, it is followed by the onset of a new dialog act and thus shares some characteristics with boundaries. For this work, we chose to put disruptions in the boundary class.

[2] There is a slight difference between the experiment and analysis sections with respect to the treatment of floor-grabbers and floor-holders (**f**, an infrequent class). In §3, such transitions are treated as boundaries, for historical reasons. Their treatment as boundaries versus nonboundaries makes little difference, however, in classifier experiments based on word recognition output, and we predict a similar lack of difference when using reference transcripts. In §4, they are treated as nonboundaries, arguably the more appropriate class, since by definition speakers intend to continue their utterance after producing these phenomena. We note also that dialog acts labeled as **z** (i.e., unlabelable because of inaudible words or other reasons) have been removed from both experiments and analyses.

Table 1: Dialog act classes and corresponding boundary and nonboundary classes. Candidate boundary locations in the examples are marked by "*"; corresponding preboundary words or word transitions are in bold face.

Abbreviation	Dialog Act	Example
Boundary		
s	statement	The new one is **better . ***
q	question	Is it almost **done ? ***
b	backchannel	**Uh-huh . * He's** done .
d	disruption	It's **my** – * **Thanks .**
Nonboundary		
n	nonboundary	The **new * one** is better .

For read data we examine data from the TDT4 English Broadcast News (BN) corpus (Strassel & Glenn 2003). The TDT4 corpus was collected by the Linguistic Data Consortium, and includes news stories from radio and television broadcasts, other electronic text, and web audio. In this study, we use a subset of read TDT4 English broadcast radio and television speech, mainly from professional news anchors. Orthographic transcriptions contain punctuation and include commercially produced transcripts for radio shows and closed captions for television programs. This style of read speech contains almost exclusively statements. The corpus is thus considered to have two dialog act boundary types: statement boundaries **s** (mapped from sentence-level punctuation marks) and nonboundaries **n** in all other locations.

Table 2 provides distributional details on the two data sets. Automatic classification experiments in §3 use training, held-out, and test sets. Feature distribution analyses in §4 use all data. Note, however, that unlike the case for automatic classification, which requires coverage of all word transitions in forced alignment output, analyses require that features be defined (not missing or undefined) for all tokens included.

As can be seen in Table 2, the different speaking styles differ significantly in mean sentence length, with sentences in meetings being only about half as long, on average, as those in broadcast news. The percentage of boundaries (relative to total boundaries) is thus higher in the spontaneous data. Meetings (and conversational speech in general) tend to contain syntactically simpler sentences and significant pronominalization. News speech is typically read from a transcript and more closely resembles written text. It contains, for example, appositions, center embeddings, and proper noun compounds, among other characteristics that contribute to longer sentences. Discourse phenomena also obviously differ across corpora, with meetings containing more turn exchanges, more incomplete sentences, and higher rates of short backchannels (such as "yeah" and "uh-huh") than speech in news broadcasts.

Table 2: Data set statistics. The BN corpus is not annotated for dialog act boundary subtypes; all boundaries are considered to be statement boundaries because classes marked by "–" have an estimated low occurrence rate.

	MRDA	BN
Automatic classification data sets		
Training set words	456,486	800,000
Test set words	87,576	82,644
Held-out set words	98,433	81,788
Vocabulary size (unique words)	11,894	21,004
Mean sentence length (words)	7.7	14.7
Distribution of dialog act boundaries	% of Total words	
Nonboundary (within dialog act)		
n nonboundary	86.55	93.20
Boundary (end of dialog act)		
s statement	8.62	6.80
q question	0.95	–
b backchannel	1.88	–
d disruption	1.99	–

2.2 Time alignments

Features are computed automatically based on time marks for words and phones. Since our interest is in inherent feature comparisons, we use forced alignments (rather than free word recognition) to avoid confounds attributable to differences in speech recognition performance rather than to the features themselves. To obtain time marks, audio waveforms were force aligned to reference orthographic transcripts using a state-of-the-art speech recognizer appropriate for each corpus (Zhu et al. 2005, Venkataraman et al. 2004). Because in this procedure words are constrained to reference words, accuracy is typically quite good. Errors can occur for specific examples, but since feature distributions are compared with other feature distributions using the same time-marking procedure, distances between the distributions overall should be largely unaffected. This consideration, coupled with the use of large amounts of data, makes it reasonable to assume that time-mark inconsistencies should show up as noise rather than bias results. The reference transcriptions of the BN data are quick transcriptions from closed captions and thus are expected to contain errors. We therefore use a flexible alignment procedure (Stolcke et al. 2006) that allows for the possibility of skipping or inserting words based on acoustic evidence.

2.3 Features

We associate each word transition with the word preceding it, as illustrated in Fig. 1. Features are extracted for each word transition, regardless of whether or not that word transition contains a pause. The features are based on any pause at the transition, on features of the word (or time window) preceding the transition, and on the word or time window following the transition. Thus, word-based features span at most two words (and any intervening pause). Longer spans could potentially yield gains, but they would also complicate matters because of the presence of short dialog acts: features that are more than one word away from a candidate boundary could pertain to a different dialog act.

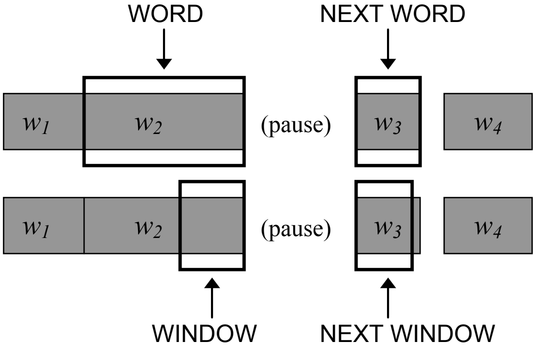

Figure 1: Illustration of feature extraction regions for the word-based (top) and window-based (bottom) features, shown for the word transition w_2, w_3, which happens to contain a pause.

2.3.1 Prosodic features

Prosodic features are extracted in the region of each word transition. The features were designed to capture breaks in temporal, intonational, and energy contours, based on the descriptive literature. Features were defined locally, since the segmentation into dialog act units is not known; one knows only that dialog act boundaries are constrained to occur at word transitions. And because of our interest in machine-based processing, features had to be extracted in a fully automatic manner, without reference to any hand labeling of prosody. Finally, for future robustness to automatic speech recognition errors, as well as for combination with lexical features, prosodic features were designed to be independent of word identities. That is, only a word's time marks, not its identity, were used for feature extraction.[3] Because of the large number of features, a general summary is provided here. Specific features of interest will be discussed in more detail in §4.

[3] Duration features, however, were normalized using information about phone-level duration statistics; strictly speaking, this uses phone identity (but not word identity) information.

Pause features included the duration of pauses as determined from recognizer forced alignment output. The pause model used by the recognizer was trained as an individual phone, which could occur optionally between words. Features included the pause duration (or 0, for no pause) at the transition, as well as that at the immediately preceding transition. Pause durations were not used for the following transition, because the presence of single-word dialog acts (such as backchannels) means that such locations may correspond to a different dialog act than that at the current word transition.

Duration features were intended to capture final lengthening before boundaries. Features included the duration of the last or maximum-duration (maximum after phone-based normalization) vowel or rhyme in the word (since lexical stress could be on other than the final syllable). We used both unnormalized and normalized versions of these features. Normalized versions were based on phone duration statistics compiled for the training data for the respective corpus.

Pitch features were based on frame-level output from a standard pitch tracker. We used an autocorrelation-based pitch tracker—the "get_f0" function in ESPS/Waves (ESPS 1993), with default parameter settings—to generate estimates of frame-level F0 (Talkin 1995). We then postprocessed the frame-level pitch to smooth out microintonation and pitch tracking errors, using median filtering followed by fitting using a piecewise linear model improved over that used in previous work (Shriberg et al. 2000, Sönmez et al. 1997). Features included the mean, maximum, minimum, first, or last value in the word or time windows shown in Fig. 1, as well as the value or sign (positive or negative) of fitted slopes in these regions. Time windows were either 200 or 500 milliseconds. Speaker-based pitch normalization used a "baseline" pitch calculated for each speaker from that talker's distribution of frame-level pitch values. The F0 distribution was modeled by three lognormal modes spaced log 2 apart in the log frequency domain. Locations of the modes were modeled with one tied parameter (μ – log 2, μ, μ + log 2), variances were scaled to be the same in the log domain, and mixture weights were estimated by an expectation maximization (EM) algorithm. Baseline pitch was estimated as the value occurring halfway between the middle and lower modes in the log domain, representing the lower end of the normal voicing mode for that speaker. The resulting measures were combined to create two types of features: those that look only at pretransition speech, and those that look at differences in pitch before and after the transition.

Energy features were based on frame-level RMS energy values from the "get_f0" function in ESPS/Waves (ESPS 1993), with default parameter settings, and were postprocessed using a piecewise linear model in a manner similar to that used for pitch regularization. Like the pitch features, energy features include mean, maximum, minimum, starting, and ending energy in the word or time windows shown in Fig. 1, as well as values and signs of fitted energy contours in these regions. Energy features were normalized based on the distribution of energy values from each speaker or channel and, like pitch features, include both pretransition features and difference features that compare energy values across the transition.

2.3.2 Lexical features

For purposes of comparison in automatic processing experiments, we also extract lexical features, which have been found to be helpful in previous work for similar tasks (Shriberg et al. 2000, Shriberg & Stolcke 2004, Liu et al. 2006, Cuendet et al. 2007b, Cuendet et al. 2007a). Lexical features are usually represented as *N*-grams of words. We represent lexical information using five *N*-gram features for each word transition, where "current" refers to the first word in the word transition:

- unigrams: {previous}, {current}, {next}
- bigrams: {current, next}
- trigram: {previous, current, next}

Because we use reference transcripts for prosodic features, for the reasons described earlier, we also use reference words for lexical features. Lexical feature results are thus optimistic compared to results using automatic speech recognition. Indeed, prior studies (Shriberg et al. 2000, Liu et al. 2006, Cuendet et al. 2007b) show that the degradation from errorful speech recognition output tends to be higher for lexical than for prosodic features for this type of task. This is particularly true for the meeting corpus, which shows a larger degradation than does BN when using recognized versus reference words.

2.3.3 Turn and overlap features

In experiments, a binary "turn" feature was included to capture locations of speaker change, rather than the more complex construct of a turn as defined in conversation analysis. The latter requires more sophisticated information than we can reliably extract automatically. For example, we do not have *a priori* knowledge of dialog act boundaries (they are, after all, what we seek to predict) or dialog act labels, so we cannot distinguish true interrupts from backchannels that do not disrupt another speaker's turn. The turn feature was necessarily computed differently for the two corpora. The meeting corpus records each speaker on an individual channel. Start and end times for talk from each speaker can thus be inferred from recognizer forced alignments on each of the individual recordings. In the broadcast news speech, however, a single channel contains multiple (generally nonoverlapping) talkers. Reference speaker information was not available for this data, so speaker change locations were estimated by an alignment between the output of an external diarization system (Wooters et al. 2004) and the reference words. Because the diarization system inserts a turn boundary at pauses greater than 500 milliseconds, the same treatment was applied to the meeting data for consistency in analyses.[4]

For analyses, particularly for conditioning of pause length distributions (§4.1), we also computed a more sophisticated feature based on overlapping speech in the region of the foreground speaker's candidate boundaries. The overlap feature asked how many other talkers (including 0) produced speech overlapping with the region extending from the onset of the foreground speaker's preboundary word through the end of the foreground speaker's postboundary word. We mapped a value of 1 or more overlapping speakers to "overlap" and 0 overlapping speakers to "no overlap."

2.4 Classifier

Dialog act segmentation can be seen as a binary classification problem, in which every word transition is to be labeled as either a dialog act boundary or a nonboundary. (More detailed models may distinguish among different types of dialog acts, such as statements versus questions.) For automatic classification experiments we use the AdaBoost algorithm (Schapire & Singer 2000), which combines weak multiple base classifiers to produce a strong classifier. In each iteration in the learning algorithm, a different distribution or weighting over the training examples is used, which gives more emphasis to examples that are often misclassified by the preceding weak classifiers. We use the BoosTexter tool (Schapire & Singer 2000); the weak learners are one-level decision trees.

[4] We note, however, that this treatment results in a higher rate of spurious turn marks in the meeting data than in news speech, because speakers in natural conversation often pause for that amount of time or longer within their turns. Since the pause feature is not affected by this treatment, but the turn feature is associated with increased false alarms, the effectiveness of the turn feature for dialog act boundary detection should be reduced for the meeting corpus (relative to what it would be were turns marked perfectly). Nevertheless, we will see later that the turn feature is still much more useful in meetings than in news speech for this task.

2.5 Metrics

We use two metrics in this study: (1) F-measure to evaluate automatic processing and (2) the Kolmogorov-Smirnov *D* statistic to quantify the difference between two feature distributions. F-measure, used in §3, is the harmonic mean of recall and precision. It is the metric traditionally used for various segmentation tasks. Note that it is an asymmetrical detection-based metric, dependent on the specific class of interest. Typically, this is the marked class (in our case, dialog act boundaries rather than nonboundaries). An F-measure of 1.0 indicates both perfect recall and perfect precision.

In the feature distribution analyses in §4, our goal is to discover which features provide best inherent separability between boundaries and nonboundaries. Since the features are not always normally distributed we use a nonparametric statistic. The Kolmogorov-Smirnov (K-S) test (Siegel & Castellan, Jr. 1988, Press et al. 1988) asks whether two data distributions differ significantly. We are interested in the value of the K-S *D* statistic, which measures the maximum difference between two cumulative distribution functions, and which can be compared directly across pairwise tests differing in sample size. The K-S statistic is

$$D = \max_{-\infty < x < \infty} | S_1(x) - S_2(x) |$$

where S_1 and S_2 are the two cumulative distribution functions to be compared: $S_i(x)$ is the percentage of the population in distribution *i* falling below *x*. It should be noted, however, that while larger *D* values reflect larger differences between two distributions, this does not guarantee that the two distributions can be separated by any specific classifier.

3. Automatic classification results

We first look at results from automatic classification using boosting with sets of features by feature type. Table 3 shows results in terms of F-measure. Chance performance is computed as follows. We assume that one has knowledge of the prior probability of a dialog act boundary in each corpus, since we know the average dialog act length in words. We compute this probability $p_t(s)$ for the training set and classify each word transition in the test set as a boundary with probability $p_t(s)$. The chance score is evaluated by computing the probability of each error and correct class and the ensuing value for the F-measure computation.

Table 3: F-measure results for automatic classification, by feature type(s) used in classifier. MRDA = meeting data, BN = TDT4 Broadcast News data.

Feature(s)	F-measure MRDA	F-measure BN
Chance	15.8	6.9
Duration	46.3	27.5
Energy	50.3	44.5
Pitch	53.8	50.5
Turn	63.9	36.7
Pause	69.5	63.7
All prosody	71.0	68.3
Lexical only	74.9	48.1
Combination (prosody + lexical)	82.0	76.9

As shown, in absolute F-measure, the meeting data has somewhat better performance than the news data on all feature types. It has particularly better performance for duration, turn, and lexical features. However, meetings also have a higher chance F-measure because of their shorter sentences. To adjust for the differences in chance performance, we should look at relative error reduction. Since the F-measure is a harmonic mean of two error types, one can compute the relative error reduction for a model with F-measure F and the associated chance performance c as

$$(1) \quad \delta = \frac{(100-c)-(100-F)}{100-c} = \frac{F-c}{100-c}$$

Results for relative error reduction are plotted in Fig. 2. A first observation about Fig. 2 is that despite the two very different speaking styles, recording conditions, speaker populations, and extremes of naturalness, relative error reduction is nearly identical for the corpora for energy, pitch, and pause features. Duration features are less useful than pitch or energy features, and pause features are the most useful individual feature type. The corpora differ in relative error reduction using duration, turn, and lexical features. Overall results when combining lexical and all prosodic features are similar.

Three of the feature types—duration, energy, and pitch—can be considered core prosodic features. It is surprising that energy and pitch features behave almost identically in degree of error reduction, given that read speech is typically considered more well-behaved prosodically. We will look more closely at these features in §4.

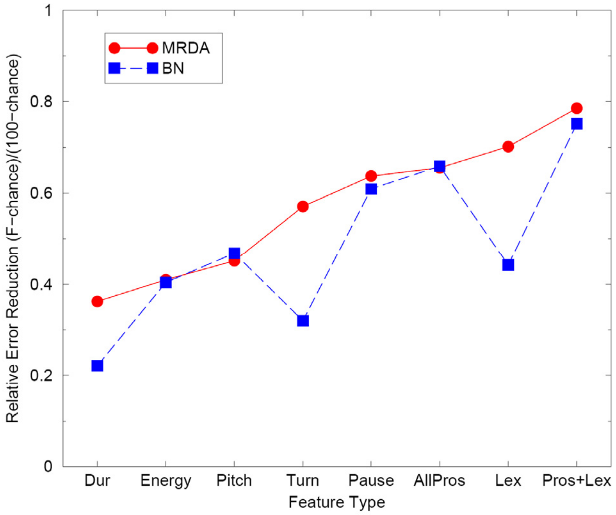

Figure 2: F-measure results by condition and features included in model. MRDA = ICSI meeting data; BN = TDT4 broadcast news data. Dur = duration features, AllPros = all prosodic features, Lex = lexical features only, Pros+Lex = combination of prosodic and lexical features.

Pause features reflect both prosodic patterns of the foreground speaker and speech activity from other talkers, particularly in meetings. It is interesting how close results are for the two speaking styles, particularly if one considers that pauses in meetings include time during which other speakers are talking, whereas pauses in news broadcasts typically do not. We examine the question of pausing further in §4.

As for the turn features (which are not prosodic *per se*), it is not surprising that they are more useful in meetings (in which there is more frequent alternation between speakers) than in read news broadcasts (in which anchors read many statements in sequence without a speaker change). It is also possible that there is some degradation for the news data from the use of turn labels from automatic diarization (since true speakers were not known), rather than from reference speaker information available via the separate recorded channels in MRDA.

Lexical features are more useful for boundary prediction in meetings than in news speech, a result consistent with previous work (Shriberg et al. 2000) that looked at spontaneous telephone speech rather than meetings. In the spontaneous speaking contexts, a small set of frequent words provides good cues to dialog act onsets. These include first person prounouns, certain fillers and discourse markers, and backchannels. In the case of backchannels, the dialog act consists of only the backchannel itself, so both the start and end of the dialog act are easy to identify. News speech, on the other hand, has far fewer of these elements, as well as significant noun compounding that can lead to phrasal ambiguity.

4. Analysis of feature distributions

The classifier experiments just described use sets of features and are affected by class priors. To understand inherent class separability by corpus and features, we need to look at individual feature distributions for boundaries and nonboundaries. We examine these distributions for the four prosodic feature types: pause, duration, pitch, and energy.

4.1 Pause features

As we just saw in Fig. 2, the most useful individual feature type for boundary classification in both corpora is pause duration. Here we examine more closely the behavior of the feature measuring pause duration at the word transition. Because a length of 0 should be interpreted as no pause, rather than as a very short pause, pauses are represented using a two-part feature: (1) presence of a pause (binary) and (2) length of the pause (if present). Furthermore, since the automatic word alignment approach used allows optional pauses between words, which can result in spurious short pauses, including those associated with stop gaps, we require a pause to have minimum duration of 50 milliseconds.

The pause feature is a rather special case, because in the meeting data it reflects both within-turn pauses and pauses that occur while a speaker does not have the floor (and presumably some other person is talking). That is, in the meeting data, since each speaker is recorded on a separate channel, pauses are simply those regions during which the foreground speaker is not talking. In the news broadcasts, speakers are recorded on a single channel, so pauses typically reflect within-speaker prosodic phrasing. Given that meetings involve more exchanges of speaker than do news broadcasts, we would expect the corpora to differ with respect to pause distributions.

Statistics on how often pauses occur at boundary versus nonboundary words are provided in Table 4.[5] Note that rates for boundaries and nonboundaries are logically independent. For the MRDA data, two versions of the statistics are computed. The “all” version counts pauses regardless of whether other talkers produce talk during the foreground speaker’s word transition. Thus, pauses in this group may be extremely long, since they count time elapsed during other talkers’ turns. In contrast, the “no overlap” version considers only those cases in which the foreground speaker is the only talker during the word transition in question. In BN, it is estimated that there is only one person speaking at a time (although there may be background speech or noise).

[5] The ratio of boundaries to total words can be computed and compared with those provided earlier. There are slight differences because the pause statistics require that the transition be defined. Undefined transitions occur when the preboundary and postboundary words are not from the same speaker, fail to align, or are at the edge of a recording.

Table 4: Rate of pauses for boundaries and nonboundaries. Pauses have a minimum duration of 50 ms. Percentage of (non)boundaries with pause is number of (non)boundaries with pause divided by total number of (non)boundaries. All = pauses regardless of other meeting participants' speech. No overlap = no overlapping speech from another participant during current speaker's word transition. BN is estimated to contain no such overlap.

	MRDA all	MRDA no overlap	BN no overlap (est.)
Total boundaries	103,839	38,917	88,608
Boundaries with pause	77,475	20,127	71,633
% Boundaries with pause	*74.6*	*51.7*	*80.8*
Total nonboundaries	666,372	543,844	1,214,394
Nonboundaries with pause	87,043	67,007	116,523
% Nonboundaries with pause	*13.1*	*12.3*	*9.6*

Two important observations can be noted from the table. First, in terms of overall rates of pauses for both boundaries and nonboundaries, the MRDA-all and BN conditions show rather similar results, with pauses in the 75-80% range for boundary words and around 10% for nonboundary words. This similarity means that from the perspective of individual channel recordings, meetings and news speech show similar pause rates, with boundaries about 7 to 8 times as likely to contain a pause as nonboundaries. The second observation is that when overlapped transitions in MRDA are removed, so that we consider only cases in which the current speaker is talking alone, there is a significant drop in the rate of pauses for boundaries—from about 75% to about 50%. That is, in meetings, only about half of nonoverlapped sentence boundaries contain a pause, a figure that may be surprising given canonical prosodic phrasing descriptions in linguistics. The fields of conversational analysis and discourse processing, however, may explain such cases via a phenomenon called "rush-through" or other mechanisms for turn retention (Schegloff 1982, Couper-Kuhlen & Selting 1996, Wennerstrom & Siegel 2003, Local & Walker 2004). Despite these statistics, pause duration is still the top feature in terms of the performance of automatic processing experiments, as we saw in Fig. 2. One explanation is that nonoverlapped transitions are actually the minority of boundaries in meetings. As can be construed from the table, only about 40% of boundaries are not overlapped.

To understand pause behavior when pauses do occur, we also look at the distribution of pause lengths at boundary versus nonboundary locations for the two data sets. In Fig. 3, distributions are normalized to unit area for comparison of inherent differences across different class sizes. Pause lengths are plotted on a log scale, since pause distributions are typically roughly lognormal across styles and languages (Campione & Véronis 2002).

Both corpora show a positive shift in pause length for boundaries as compared with nonboundaries. The BN data shows basically a shift, with otherwise roughly similar curves for boundaries and nonboundaries. The MRDA data, however, shows some interesting behaviors. When overlapped transitions are removed, the curve for MRDA boundaries looks quite similar to that for BN boundaries, after a positive shift. Thus, pause durations at boundaries are generally longer in meetings than in read speech but follow an otherwise similar distribution. When overlapped transitions are included, meeting boundaries show a striking positive tail, suggesting at least two underlying distributions, one of which must correspond to time during which other talkers are speaking. Further analysis by dialog act type reveals that the longer-duration pauses are associated mainly with boundaries after backchannels. That is, pauses after backchannels are long, because they correspond to the time during which another speaker has the floor. A final interesting observation concerns pause lengths at nonboundaries. In this case, the shape of the MRDA distribution (similar for overlapped and nonoverlapped transitions) differs from the shapes of the BN distributions and from the MRDA nonoverlapped boundary distribution. More specifically, it shows a broader and more positive range of values. We can hypothesize that the longer pauses in these nonboundary cases may correspond to hesitation pauses, which are rare in read news speech.

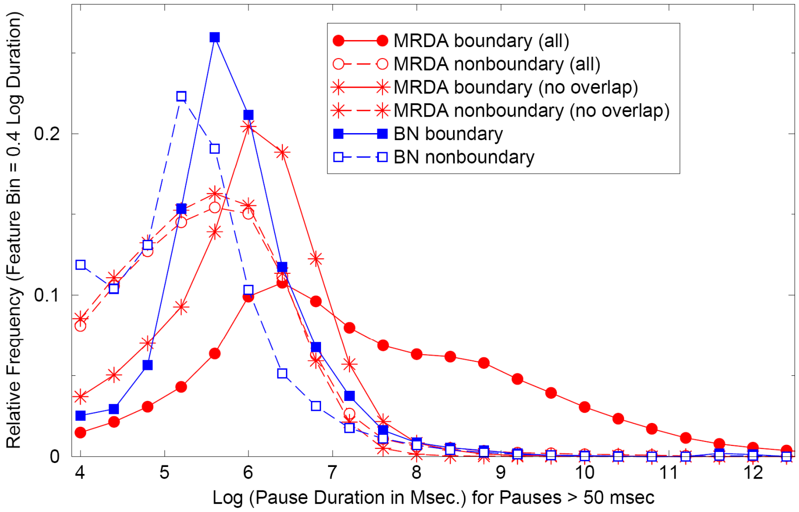

Figure 3: Distribution of pause lengths at word transitions that contain a pause (of at least 50 milliseconds), for both corpora. Length is plotted on a log scale; all distributions are normalized to unit area. MRDA is plotted in two ways: all transitions, and transitions during which no other meeting participants are speaking.

These findings on pauses show some similarities across corpora (rate of pause presence overall, shape of pause length distributions for nonoverlapped boundaries) but also differences (pause presence rates when overlaps are removed, shape of nonboundary distributions). Because pause features are typically the most powerful features for automatic classification, as seen in §3, the differences, as well as differences in corpus priors for boundaries and nonboundaries, are likely to affect results for all remaining features in machine learning experiments.

4.2 Duration features

We look first at features based on phone durations. Of the eight available features (based on rhymes versus vowels, maximum versus last syllables in words, and different normalizations) the best separation between boundary and nonboundary distributions is given by the feature RHYME_NORM_DUR_PH_MAX, which computes the mean of *Z*-score-normalized phone durations in each rhyme and selects the maximum value over all rhymes in the word. For both corpora, more separation is provided by rhyme-based features than by vowel-based features and by maximum-normalized-duration rhymes than by ending rhymes or longest rhymes in a word.

Fig. 4 shows distributions for RHYME NORM DUR PH MAX for boundary and nonboundary words in both corpora. In raw terms, news speech has slightly longer rhyme durations than meeting speech. Since durations in the figure are normalized based on the statistics of phones in the training data for each corpus, respectively, data for the majority class (i.e., nonboundaries) should look fairly similar across corpora. As shown, this is the case—although the meeting speech has a tighter distribution for nonboundaries than does the read data. The news speech shows a tendency for duration-increased nonboundaries at normalized durations between roughly 0 and 1. Listening analyses suggest that this behavior reflects a tendency for news anchors to produce frequent prominent syllables, including on nonboundary (and even function) words, perhaps to maintain listener attention.

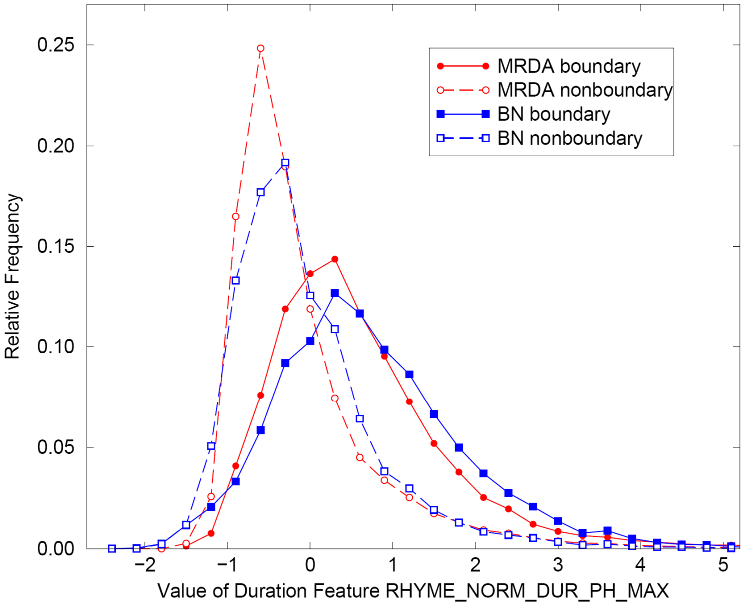

Figure 4: Distribution of the value of RHYME_NORM_DUR_PH_MAX, the maximum normalized rhyme duration in the word. Normalized rhyme duration is computed as the average of mean-and variance-normalized phone durations in the rhyme.

The most important and interesting observation from Fig. 4 is that the positive shift from nonboundary to boundary distributions is similar in both corpora. This is not simply an epiphenomenon of phone-duration normalization. Boundaries could have shown different variance versus shift patterns in the two corpora. Indeed, we might have expected spontaneous speech to show less preboundary lengthening and more variability in lengthening, but there is no evidence from the figure that this is the case. Instead, it appears that after adjusting for the difference in speaking rate via phone-based normalization, speakers use durational lengthening before boundaries to roughly the same degree in spontaneous speech as in read speech.

4.3 Pitch features

We examined inherent separation for 37 pitch features. Features were based on differences across the boundary, pitch relative to baseline estimated pitch, and pitch slopes based on lognormal fits. We look first at the top three for each corpus, based on the value of the K-S *D* statistic described in §2.5. This measures the difference between the distributions for the boundary versus nonboundary class within a corpus, by feature. Results are given in Table 5.

Table 5: Top three (out of 37 analyzed) pitch features for each corpus, based on K-S statistic for comparison of distributions for boundaries versus nonboundaries. Feature types: pitch diff = comparison of pitch level before and after boundary; pitch baseln = comparison of pitch level in word to estimated pitch baseline for speaker.

Corpus	Feature type	Feature	K-S *D*	Prob(*D*)
MRDA	pitch diff	WRD_F0K_DIFF_ENDBEG	0.303	0
	pitch diff	F0K_WRD_DIFF_LOLO_N	0.277	0
	pitch diff	F0K_20_20_WIN_DIFF_LOLO_N	0.265	0
BN	pitch diff	WRD_F0K_DIFF_ENDBEG	0.441	0
	pitch baseln	F0K_DIFF_LAST_KBASELN	0.433	0
	pitch baseln	F0K_LR_LAST_KBASELN	0.432	0

Rather remarkably, the feature that shows the most distributional differences between boundaries and nonboundaries is the same for both corpora: WRD_F0K_DIFF_ENDBEG. This feature is computed as the log ratio of the last good prestylized pitch in the last word before the boundary and the first good stylized pitch in the first word after the boundary. It is intended to capture pitch reset (from lower to higher pitch) in the case of boundaries, and thus should have a lower value for boundaries than for nonboundaries. Data is plotted in Fig. 5.

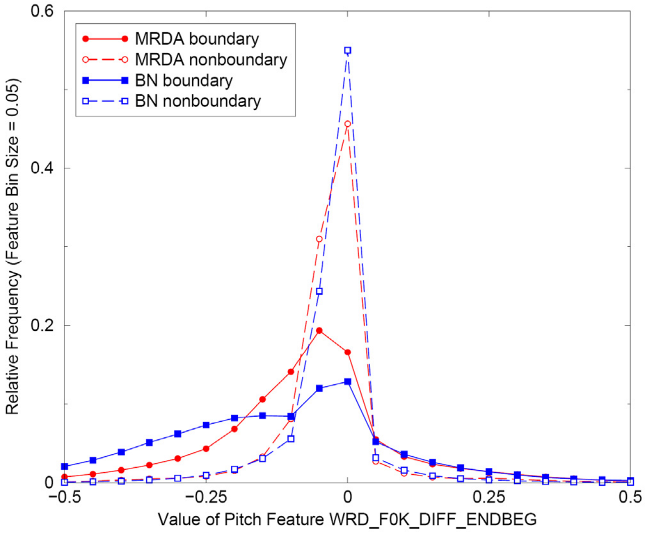

Figure 5: Distribution of the value of WRD F0K DIFF ENDBEG, the (natural) log ratio of the value of the last (fitted) pitch in the word to that of the first (fitted) pitch in the next word.

As predicted, values for the boundaries are lower than for nonboundaries. Four additional observations can be made from the figure. First, the nonboundary values are nearly identical for the two corpora; this is not a function of the feature computation, but rather reflects that nonboundary transitions in pitch are basically the same in read speech as in meeting speech—despite the different styles and also the different speaker populations in the two corpora. In general, this pattern of similar nonboundary distributions holds for many of the pitch features examined. Second, the boundary distribution for BN is more negative than that for MRDA, meaning the speakers in news broadcasts create larger pitch resets at dialog act boundaries than do meeting participants. Because of the increased value of baseline-related features in BN (discussed next), this is most likely attributable not only to higher starting pitch following a boundary but also to a drop to a lower pitch preceding the boundary. Third, both boundary distributions, and particularly the BN distribution, show a knee at roughly -0.1. The knee also occurs for some of the component dialog act boundaries and thus does not appear to be attributable to composition of different dialog-act-specific distributions. Although it requires further study, one possibility is that the knee reflects two categories of pitch distributions—one in which no pause is present (and thus reset is more constrained) and one in which a pause separates the pre- and post-transition words (and for which pitch jumps can be much larger). Finally, while the trends are similar across corpora, BN has overall higher discriminability. We will see why this is the case when we look at preboundary features, in Fig. 6.

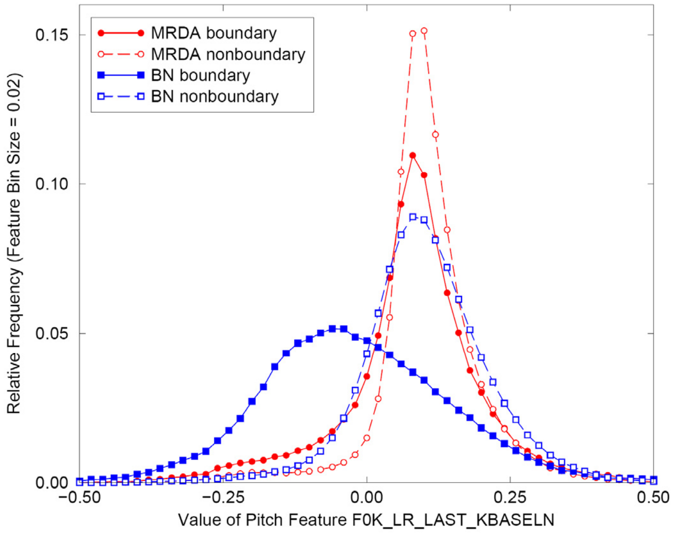

Figure 6: Distribution of the value of F0K LR LAST KBASELN, the (natural) log ratio of the value of the last (fitted) pitch in the word to that speaker's estimated baseline pitch value.

Looking back at Table 5, it can also be seen that after the shared best feature, the types of features in the two lists diverge. Features involving a speaker's baseline pitch are relatively more useful in news speech than in meeting speech for the boundary versus nonboundary distinction. The baseline features express the pitch range of the word in question by comparing it with a reference or baseline pitch for the speaker (Sönmez et al. 1997). This holds as a general trend when all 37 features are included: while both corpora have strong cross-boundary pitch cues, only read speech shows strong preboundary pitch cues.

This difference can be understood by looking at distributions for the preboundary pitch features. We choose the feature F0K_LR_LAST_KBASELN as an example, but results look quite similar for other features comparing the pitch of this word from other locations (e.g., mean, max, min in the word) with the baseline value. The feature F0K_LR_LAST_KBASELN is computed as the log ratio between the last good pitch value in the word preceding a boundary and the speaker's estimated baseline pitch.

As shown in Fig. 6, there is minimal separation between boundary and nonboundary words for the spontaneous speech, while read speech shows a negative shift for the boundary class. In meetings, data for both classes centers at about the same positive value as for the nonboundary class in news speech. These trends imply that before dialog act boundaries, speakers in read speech drop pitch to a value that is close to their baseline (on average, to slightly below it). In meetings, this preboundary pitch drop is much less common (only a small percentage of cases display more negative values than does the nonboundary class).

One reason for the lower prevalence of preboundary pitch drop in meetings may be that in spontaneous speech, pitch varies for reasons beyond phrasing, including paralinguistic factors such as affective state or emotion. Another reason may be that lexical information is a relatively stronger cue to dialog act boundaries in meetings than in news speech, as seen earlier. Phenomena such as first person pronouns, fillers, discourse markers, and backchannels all provide useful cues to dialog act onsets in meetings; such elements are much less common in news speech. Thus, there may be trading relationships between the lexical and prosodic cues. Finally, because dialog acts are, on average, shorter and less complex in meetings than in the news data, there may be less need to mark dialog act boundaries prosodically.

Preboundary pitch features are, however, among the best features in meetings for distinguishing different dialog acts—particularly for detecting questions. As can be seen in Fig. 7, questions show a shift to the right of the nonboundary distribution (rather than to the left, as in statement boundaries), reflecting the tendency of questions to end in higher pitch values.

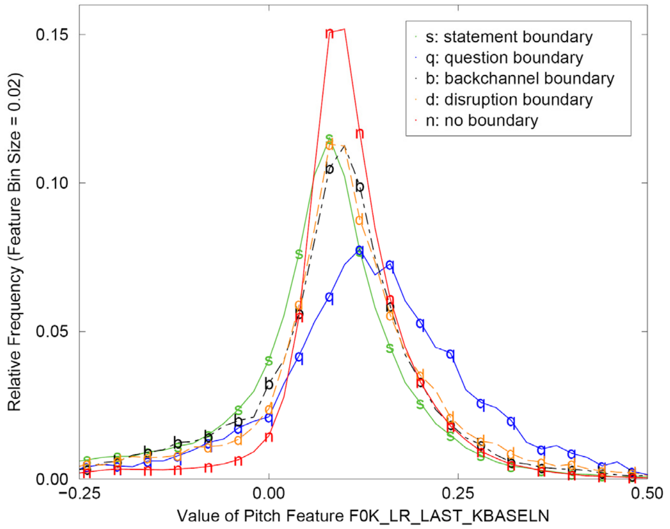

Figure 7: Dialog-act-specific distributions of the value of F0K_LR_LAST_KBASELN, the (natural) log ratio of the value of the last (fitted) pitch in the word to that speaker's estimated baseline pitch value, in the MRDA corpus. Dialog act labels are marked at every other bin.

4.4 Energy features

Out of 21 analyzed energy features, the same features appear in the top three for both corpora, as shown in Table 6. Unlike the case for pitch, good energy features require a comparison of pre- and post-boundary values in both corpora. Preboundary energy features (i.e., features using only one extraction point for feature computation) were weak in both the news and meeting data. This may reflect both the inherent nature of energy and the difficulty in normalizing energy effectively enough across different speakers and recordings to enable use of only a single extraction region.

Table 6: Top three energy features for each corpus, based on K-S statistic for comparison of distributions for boundaries versus nonboundaries. Feature types: energy diff = comparison of energy level before and after boundary; energy slope diff = difference in energy slope before and after boundary.

Corpus	Feature type	Feature	K-S *D*	Prob(*D*)
MRDA	energy diff	ENERGY_20_20_WIN_DIFF-HIHI_N	0.295	0
	energy diff	WRD_ENERGY_DIFF_ENDBEG	0.275	0
	energy slope diff	Slope_ENERGY_DIFF	0.272	0
BN	energy diff	WRD_ENERGY_DIFF_ENDBEG	0.398	0
	energy diff	ENERGY_20_20_WIN_DIFF_HIHI_N	0.306	0
	energy slope diff	Slope_ENERGY_DIFF	0.240	0

We look at the feature WRD_ENERGY_DIFF_ENDBEG, which has the best combined *D* value. Interestingly, this feature is the energy version of the pitch feature we saw earlier, which compared the pitch value at the end of the word in question with that at the start of the next word. For the present feature, fitted energy is used rather than fitted pitch. The distribution for boundaries and nonboundaries is shown in Fig. 8.

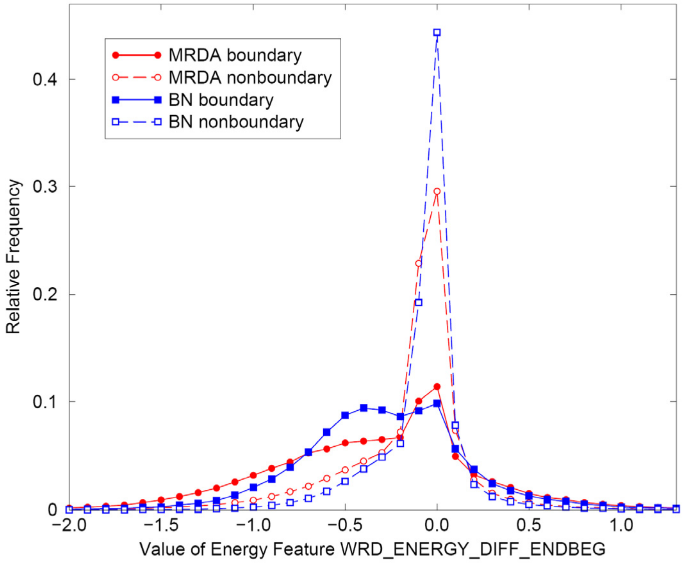

Figure 8: Distribution of the value of WRD_ENERGY_DIFF_ENDBEG, the log ratio of the last good fitted energy value in the word to that of the first good energy value in the next word.

The energy distributions look similar to those seen for the corresponding pitch feature. Knees in the distributions may reflect whether or not a pause is present at the word transition. As can be seen in Fig. 9, all dialog acts show a similar knee, except backchannels. This would make sense given the pause hypothesis, since backchannels have a high rate of following pauses.

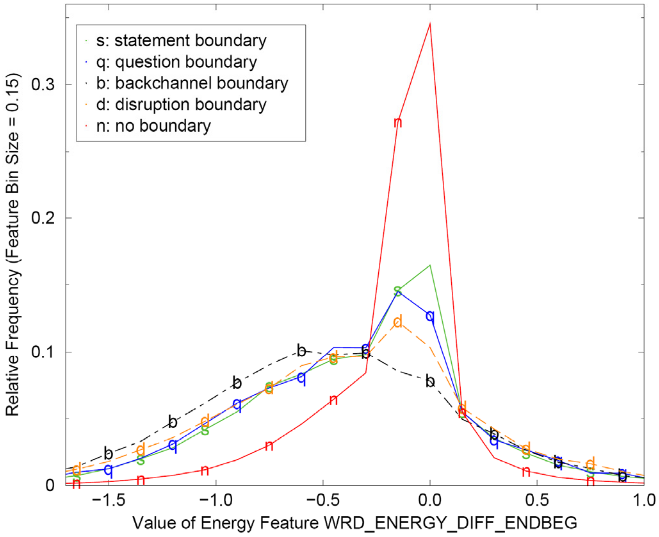

Figure 9: Dialog-act-specific distributions of the value of WRD_ENERGY_DIFF_ ENDBEG, the log ratio of the last good fitted energy value in the word to that of the first good energy value in the next word, in the MRDA corpus. Dialog act labels are marked at every other data point.

5. Summary and discussion

Results using feature groups in automatic classification experiments show that after adjusting for a measure of chance performance appropriate for F-measure analyses, meeting speech and news speech show similar dialog act segmentation performance for pitch, energy, and pause features. Perhaps counterintuitively at first, meetings show better performance for duration and lexical features. Lexical features are most likely more powerful in meetings because of the frequent occurrence of fillers, discourse markers, and first person pronouns at dialog act onsets. Turn features also differ, but in an expected manner given the different discourse contexts.

In analyses of feature distributions, it was seen that despite differences in turn-taking between the corpora, overall rates of the presence of a pause (of at least 50 milliseconds) were not that different when using very simple pause extraction definitions. Distributions of pause lengths given the presence of a pause differed, with longer pauses in MRDA, as expected, but the shapes and locations of the distributions suggest that a simple scaling could be used for adaptation. A small difference between the corpora in distribution shape for boundaries was explained by the distribution for backchannels, which show particularly long following pauses (consistent with their function of encouraging another speaker to continue talking).

Differences in the utility of duration features in the automatic experiments appear to be attributable, at least in part, to the tendency of news anchors to use longer durations for some nonboundaries, perhaps to keep the attention of listeners. When duration distributions are normalized for speaking rate, however, it appears that speakers in meetings use about the same amount of relative lengthening for boundaries as do speakers in news broadcasts.

Pitch and energy feature analyses showed remarkable similarities for the two corpora, both in terms of which features provided best inherent boundary/nonboundary separation within a corpus and in terms of the similarity of the feature distributions themselves. One difference between the two styles is that preboundary pitch drop appears to be more systematic in news speech than in meeting speech. Both styles, however, make good use of features that compare pitch across the transition, and these difference features also are the most robust for energy. Preboundary pitch is, however, important for distinguishing questions from other dialog acts in meeting speech (questions did not occur frequently enough in news speech to assess results).

The current study examined only two speaking styles, and clearly additional corpora and genres should be investigated to better understand prosodic consistencies for this task. Nevertheless, given the very different styles examined, results offer promise for improved approaches to cross-genre prosodic feature modeling for this task. Previous work in this domain has shown that if two speaking styles share characteristics, one can perform automatic adaptation from one style to another, to improve segmentation performance. An example has been shown recently for meeting data in a study of adaption using conversational telephone speech (Cuendet et al. 2006). Results such as those shown here could be used to help in selecting training data to better match characteristics of a test set. Measures of feature divergence could also be used more generally for feature selection.

But more importantly, we propose that feature distributions themselves could be shared. Classifiers will generalize to new data to the extent that feature distributions are invariant to genres, speakers, recordings, and other sources of variability. Features can possibly be made invariant by suitable normalizations, and normalization is facilitated by distributions that differ in simple ways—e.g., by simple scaling and shifts. A condition for effective normalization is that the parameters of the feature transformation can be estimated from data. For example, if the class-conditional distributions are unimodal and an optimal decision threshold is fixed relative to the overall distribution mean, it should be possible to estimate the decision threshold from the test data. An alternative approach would be to map the features from different domains to a common feature space, using techniques developed for channel compensation in speaker recognition (Reynolds 2003).

To benefit from such normalizations, one would also need to consider differences in class priors, which obviously affect the optimal classifier decision. Class priors can be gleaned from matched training data (using transcripts, with no prosodic feature extraction or modeling necessary). But in the absence of matched training data, this parameter could also be estimated in unsupervised fashion from test data. This could be achieved by running a preliminary classifier on a body of test data, using the outputs to update estimates of class priors, and reclassifying until estimates stabilize.

If features that differ across genres could be normalized so as to make their distributions similar for all genres, then classifiers could be trained on data irrespective of style and would generalize to new data mismatched for genre. An alternative approach could be to train the classifier to perform the normalization as part of the training process. In principle, this could be achieved by giving the data source as an additional input feature, so that the classifier could learn to adjust feature values to be usable for the same task across different genres.

6. Conclusions

Feature selection experiments using machine learning approaches yield different prosodic feature sets for dialog act segmentation, depending on the corpus. Yet we have shown herein that inherent prosodic feature distributions are remarkably similar across styles. This suggests, rather surprisingly, that prosodic marking of dialog boundaries in spontaneous speech is much like that in read speech. Differences in feature selection are likely a consequence of differences in class priors and in pause thresholds across corpora, which then affect usage of remaining features in automatic classification approaches. Clearly, other features, and especially other genres of speech, need to be investigated before definitive conclusions about the genre specificity of prosodic patterns can be drawn. In future work it will be worth investigating techniques for feature normalization and prior calibration tailored to the distributions studied here, with the goal of improving the robustness and generalizability of automatic sentence segmentation algorithms across speaking styles.

References

Elizabeth Shriberg
Speech Technology & Research Laboratory
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025, USA
ees@speech.sri.com