Prosody-Based Automatic Segmentation of Speech into Sentences and Topics

Elizabeth Shriberg   Andreas Stolcke
Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, U.S.A.
{ees,stolcke}@speech.sri.com

Dilek Hakkani-Tür   Gökhan Tür
Department of Computer Engineering, Bilkent University
Ankara, 06533, Turkey
{hakkani,tur}@cs.bilkent.edu.tr

Abstract

A crucial step in processing speech audio data for information extraction, topic detection, or browsing/playback is to segment the input into sentence and topic units. Speech segmentation is challenging, since the cues typically present for segmenting text (headers, paragraphs, punctuation) are absent in spoken language. We investigate the use of prosody (information gleaned from the timing and melody of speech) for these tasks. Using decision tree and hidden Markov modeling techniques, we combine prosodic cues with word-based approaches, and evaluate performance on two speech corpora, Broadcast News and Switchboard. Results show that the prosodic model alone performs on par with, or better than, word-based statistical language models—for both true and automatically recognized words in news speech. The prosodic model achieves comparable performance with significantly less training data, and requires no hand-labeling of prosodic events. Across tasks and corpora, we obtain a significant improvement over word-only models using a probabilistic combination of prosodic and lexical information. Inspection reveals that the prosodic models capture language-independent boundary indicators described in the literature. Finally, cue usage is task and corpus dependent. For example, pause and pitch features are highly informative for segmenting news speech, whereas pause, duration and word-based cues dominate for natural conversation.

1

Zusammenfassung

Ein wesentlicher Schritt in der Sprachverarbeitung zum Zweck der Informationsextrahierung, Themenklassifizierung oder Wiedergabe ist die Segmentierung in thematische und Satzeinheiten. Sprachsegmentierung ist schwierig, da die Hinweise, die dafür gewöhnlich in Texten vorzufinden sind (Überschriften, Absätze, Interpunktion), in gesprochener Sprache fehlen. Wir untersuchen die Benutzung von Prosodie (Timing und Melodie der Sprache) zu diesem Zweck. Mithilfe von Entscheidungsbäumen und Hidden-Markov-Modellen kombinieren wir prosodische und wortbasierte Informationen, und prüfen unsere Verfahren anhand von zwei Sprachkorpora, Broadcast News und Switchboard. Sowohl bei korrekten, als auch bei automatisch erkannten Worttranskriptionen von Broadcast News zeigen unsere Ergebnisse, daß Prosodiemodelle alleine eine gleichgute oder bessere Leistung als die wortbasieren statistischen Sprachmodelle erbringen. Dabei erzielt das Prosodiemodell eine vergleichbare Leistung mit wesentlich weniger Trainingsdaten und bedarf keines manuellen Transkribierens prosodischer Eigenschaften. Für beide Segmentierungsarten und Korpora erzielen wir eine signifikante Verbesserung gegenüber rein wortbasierten Modellen, indem wir prosodische und lexikalische Informationsquellen probabilistisch kombinieren. Eine Untersuchung der Prosodiemodelle zeigt, daß diese auf sprachunabhängige, in der Literatur beschriebene Segmentierungsmerkmale ansprechen. Die Auswahl der Merkmale hängt wesentlich von Segmentierungstyp und Korpus ab. Zum Beispiel sind Pausen und F0-Merkmale vor allem für Nachrichtensprache informativ, während zeitdauer- und wortbasierte Merkmale in natürlichen Gesprächen dominieren.

Resumé

Une étape cruciale dans le traitement de la parole pour l'extraction d'information, la détection du sujet de conversation et la navigation est la segmentation du discours. Celle-ci est difficile car les indices aidant à segmenter un texte (en-têtes, paragraphes, ponctuation) n'apparaissent pas dans le language parlé. Nous étudions l'usage de la prosodie (l'information extraite du rythme et de la mélodie de la parole) à cet effet. A l'aide d'arbres de décision et de chaînes de Markov cachées, nous combinons les indices prosodiques avec le modèle du langage. Nous evaluons notre algorithme sur deux corpora, Broadcast News et Switchboard. Nos résultats indiquent que le modèle prosodique est équivalent ou supérieur au modèle du langage, et qu'il requiert moins de données d'entraînement. Il ne nécessite pas d'annotations manuelles de la prosodie. De plus, nous obtenons un gain significatif en combinant de manière probabiliste l'information prosodique et lexicale, et ce pour différents corpora et applications. Une inspection plus détaillée des résultats révèle que les modèles prosodiques identifient les indicateurs de début et de fin de segments, tel que décrit dans la litterature. Finalement, l'usage des indices prosodiques dépend de l'application et du corpus. Par exemple, le ton s'avère extrèmement utile pour la segmentation des bulletins télévisés, alors que les caracteristiques de durée et celles extraites du modèle du langage servent davantage pour la segmentation de conversations naturelles.

3

1 Introduction

1.1 Why process audio data?

Extracting information from audio data allows examination of a much wider range of data sources than does text alone. Many sources (e.g., interviews, conversations, news broadcasts) are available only in audio form. Furthermore, audio data is often a much richer source than text alone, especially if the data was originally meant to be *heard* rather than read (e.g., news broadcasts).

1.2 Why automatic segmentation?

Past automatic information extraction systems have depended mostly on lexical information for segmentation (Kubala et al., 1998; Allan et al., 1998; Hearst, 1997; Kozima, 1993; Yamron et al., 1998, among others). A problem for the text-based approach, when applied to speech input, is the lack of typographic cues (such as headers, paragraphs, sentence punctuation, and capitalization) in continuous speech.

A crucial step toward robust information extraction from speech is the automatic determination of topic, sentence, and phrase boundaries. Such locations are overt in text (via punctuation, capitalization, formatting) but are absent or "hidden" in speech output. Topic boundaries are an important prerequisite for topic detection, topic tracking, and summarization. They are further helpful for constraining other tasks such as coreference resolution (e.g., since anaphoric references do not cross topic boundaries). Finding sentence boundaries is a necessary first step for topic segmentation. It is also necessary to break up long stretches of audio data prior to parsing. In addition, modeling of sentence boundaries can benefit named entity extraction from automatic speech recognition (ASR) output, for example by preventing proper nouns spanning a sentence boundary from being grouped together.

1.3 Why use prosody?

When spoken language is converted via ASR to a simple stream of words, the timing and pitch patterns are lost. Such patterns (and other related aspects that are independent of the words) are known as speech *prosody*. In all languages, prosody is used to convey structural, semantic, and functional information.

Prosodic cues are known to be relevant to discourse structure across languages (e.g., Vaissière, 1983) and can therefore be expected to play an important role in various information extraction tasks. Analyses of read or spontaneous monologues in linguistics and related fields have shown that information units, such as sentences and paragraphs, are often demarcated prosodically. In English and related languages, such prosodic indicators include pausing, changes in pitch range and amplitude, global pitch declination, melody and boundary tone distribution, and speaking rate variation. For example, both sentence boundaries and paragraph or topic boundaries are often marked by some combination of a long pause, a preceding final low boundary tone, and a pitch range reset, among other features (Lehiste, 1979, 1980; Brown et al., 1980; Bruce, 1982; Thorsen, 1985; Silverman, 1987; Grosz and Hirschberg, 1992; Sluijter and Terken, 1994; Swerts and Geluykens, 1994; Koopmans-van Beinum and van Donzel, 1996; Hirschberg and Nakatani, 1996; Nakajima and Tsukada, 1997; Swerts, 1997; Swerts and Ostendorf, 1997).

Furthermore, prosodic cues by their nature are relatively unaffected by word identity, and should therefore improve the robustness of lexical information extraction methods based on ASR output. This may be particularly important for spontaneous human-human conversation since ASR word error rates remain much higher for these corpora than for read, constrained, or computer-directed speech (National Institute for Standards and Technology, 1999).

A related reason to use prosodic information is that certain prosodic features can be computed even in the absence of availability of ASR, for example, for a new language where one may not have a dictionary available. Here they could be applied for instance for audio browsing and playback, or to cut waveforms prior to recognition to limit audio segments to durations feasible for decoding.

Furthermore, unlike spectral features, some prosodic features (e.g., duration and intonation patterns) are largely invariant to changes in channel characteristics (to the extent that they can be adequately extracted from the signal). Thus, the research results are independent of characteristics of the communica-

4

tion channel, implying that the benefits of prosody are significant across multiple applications.

Finally, prosodic feature extraction can be achieved with minimal additional computational load and no additional training data; results can be integrated directly with existing conventional ASR language and acoustic models. Thus, performance gains can be evaluated quickly and cheaply, without requiring additional infrastructure.

1.4   This study

Past studies involving prosodic information have generally relied on hand-coded cues (an exception is Hirschberg and Nakatani, 1996). We believe the present work to be the first that combines fully automatic extraction of both lexical and prosodic information for speech segmentation. Our general framework for combining lexical and prosodic cues for tagging speech with various kinds of *hidden* structural information is a further development of earlier work on detecting sentence boundaries and disfluencies in spontaneous speech (Shriberg et al., 1997; Stolcke et al., 1998; Hakkani-Tür et al., 1999; Stolcke et al., 1999; Tür et al., 2000) and on detecting topic boundaries in Broadcast News (Hakkani-Tür et al., 1999; Stolcke et al., 1999; Tür et al., 2000). In previous work we provided only a high-level summary of the prosody modeling, focusing instead on detailing the language modeling and model combination.

In this paper we describe the prosodic modeling in detail. In addition we include, for the first time, controlled comparisons for speech data from two corpora differing greatly in style: Broadcast News (Graff, 1997) and Switchboard (Godfrey et al., 1992). The two corpora are compared directly on the task of sentence segmentation, and the two tasks (sentence and topic segmentation) are compared for the Broadcast News data. Throughout, our paradigm holds the candidate features for prosodic modeling constant across tasks and corpora. That is, we created parallel prosodic databases for both corpora, and used the same machine learning approach for prosodic modeling in all cases. We look at results for both true words, and words as hypothesized by a speech recognizer. Both conditions provide informative data points. True words reflect the inherent additional value of prosodic information above and beyond perfect word information. Using recognized words allows comparison of

degradation of the prosodic model to that of a language model, and also allows us to assess realistic performance of the prosodic model when word boundary information must be extracted based on incorrect hypotheses rather than forced alignments.

Section 2 describes the methodology, including the prosodic modeling using decision trees, the language modeling, the model combination approaches, and the data sets. The prosodic modeling section is particularly detailed, outlining the motivation for each of the prosodic features and specifying their extraction, computation, and normalization. Section 3 discusses results for each of our three tasks: sentence segmentation for Broadcast News, sentence segmentation for Switchboard, and topic segmentation for Broadcast News. For each task, we examine results from combining the prosodic information with language model information, using both transcribed and recognized words. We focus on overall performance, and on analysis of which prosodic features prove most useful for each task. The section closes with a general discussion of cross-task comparisons, and issues for further work. Finally, in Section 4 we summarize main insights gained from the study, concluding with points on the general relevance of prosody for automatic segmentation of spoken audio.

2   Method

2.1   Prosodic modeling

2.1.1   Feature extraction regions

In all cases we used only very local features, for practical reasons (simplicity, computational constraints, extension to other tasks), although in principle one could look at longer regions. As shown in Fig. 1, for each inter-word boundary, we looked at prosodic features of the word immediately preceding and following the boundary, or alternatively within a window of 20 frames (200 ms, a value empirically optimized for this work) before and after the boundary. In boundaries containing a pause, the window extended backward from the pause start, and forward from the pause end. (Of course, it is conceivable that a more effective region could be based on information about syllables and stress patterns, for example, extending backward and forward until a stressed syllable is reached. However, the recognizer used did not model stress, so we
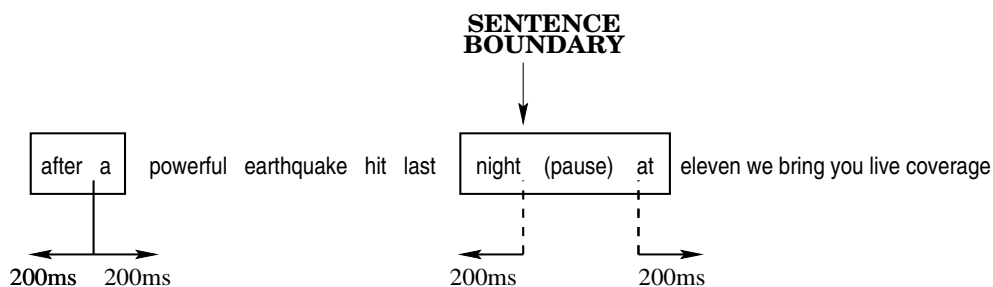
5

Fig. 1: Feature extraction regions for each inter-word boundary

preferred the simpler, word-based criterion used here.)

We extracted prosodic features reflecting pause durations, phone durations, pitch information, and voice quality information. Pause features were extracted at the inter-word boundaries. Duration, F0, and voice quality features were extracted mainly from the word or window *preceding* the boundary (which was found to carry more prosodic information for these tasks than the speech *following* the boundary; Shriberg et al., 1997). We also included pitch-related features reflecting the difference in pitch range *across* the boundary.

In addition, we included nonprosodic features that are inherently related to the prosodic features, for example, features that make a prosodic feature undefined (such as speaker turn boundaries) or that would show up if we had not normalized appropriately (such as gender, in the case of F0). This allowed us both to better understand feature interactions, and to check for appropriateness of normalization schemes.

We chose not to use amplitude- or energy-based features, since previous work showed these features to be both less reliable than and largely redundant with duration and pitch features. A main reason for the lack of robustness of the energy cues was the high degree of channel variability in both corpora examined, even after application of various normalization techniques based on the signal-to-noise ratio distribution characteristics of, for example, a conversation side (the speech recorded from one speaker in the two-party conversation) in Switchboard. Exploratory work showed that energy measures can correlate with shows (news programs in the Broadcast News corpus), speakers, and so forth, rather than with the structural locations in which we were interested. Duration and pitch, on the other hand, are relatively invariant to channel effects (to the extent that they can be adequately extracted).

In training, word boundaries were obtained from recognizer forced alignments. In testing on recognized words, we used alignments for the 1-best recognition hypothesis. Note that this results in a mismatch between train and test data for the case of testing on recognized words, that works *against* us. That is, the prosodic models are trained on better alignments than can be expected in testing; thus, the features selected may be suboptimal in the less robust situation of recognized words. Therefore, we expect that any benefit from the present, suboptimal approach would be only enhanced if the prosodic models were based on recognizer alignments in training as well.

2.1.2   Features

We included features that, based on the descriptive literature, should reflect breaks in the temporal and intonational contour. We developed versions of such features that could be defined at each inter-word boundary, and that could be extracted by completely automatic means, without human labeling. Furthermore, the features were designed to be independent of word identities, for robustness to imperfect recognizer output.

We began with a set of over 100 features, which, after initial investigations, was pared down to a smaller set by eliminating features that were clearly not at all useful (based on decision tree experiments; see also Section 2.1.4). The resulting set of features is described below. Features are grouped into broad

feature classes based on the kinds of measurements involved, and the type of prosodic behavior they were designed to capture.

2.1.2.1 Pause features.

Important cues to boundaries between semantic units, such as sentences or topics, are breaks in prosodic continuity, including pauses. We extracted pause duration at each boundary based on recognizer output. The pause model used by the recognizer was trained as an individual phone, which during training could occur optionally between words. In the case of no pause at the boundary, this pause duration feature was output as 0.

We also included the duration of the pause preceding the word before the boundary, to reflect whether speech right before the boundary was just starting up or continuous from previous speech. Most inter-word locations contained no pause, and were labeled as zero length. We did not need to distinguish between actual pauses and the short segmental-related pauses (e.g., stop closures) inserted by the speech recognizer, since models easily learned to distinguish the cases based on duration.

We investigated both raw durations and durations normalized for pause duration distributions from the particular speaker. Our models selected the unnormalized feature over the normalized version, possibly because of a lack of sufficient pause data per speaker. The unnormalized measure was apparently sufficient to capture the gross differences in pause duration distributions that separate boundary from nonboundary locations, despite speaker variation within both categories.

For the Broadcast News data, which contained mainly monologues and which was recorded on a single channel, pause durations were undefined at speaker changes. For the Switchboard data there was significant speaker overlap, and a high rate of backchannels (such as "uh-huh") that were uttered by a listener during the speaker's turn. Some of these cases were associated with simultaneous speaker pausing and listener backchanneling. Because the pauses here did not constitute real turn boundaries, and because the Switchboard conversations were recorded on separate channels, we included such speaker pauses in the pause duration measure (i.e., even though a backchannel was uttered on the other channel).

2.1.2.2 Phone and rhyme duration features.

Another well-known cue to boundaries in speech is a slowing down toward the ends of units, or preboundary lengthening. Preboundary lengthening typically affects the nucleus and coda of syllables, so we included measures here that reflected duration characteristics of the last rhyme (nucleus plus coda) of the syllable preceding the boundary.

Each phone in the rhyme was normalized for inherent duration as follows

$$\sum_i \frac{phone\_dur_i - mean\_phone\_dur_i}{std\_dev\_phone\_dur_i} \quad (1)$$

where $mean\_phone\_dur_i$ and $std\_dev\_phone\_dur_i$ are the mean and standard deviation of the current phone over all shows or conversations in the training data.[1] Rhyme features included the average normalized phone duration in the rhyme, computed by dividing the measure in Eq. (1) by the number of phones in the rhyme, as well as a variety of other methods for normalization. To roughly capture lengthening of prefinal syllables in a multisyllabic word, we also recorded the longest normalized phone, as well as the longest normalized vowel, found in the preboundary word.[2]

We distinguished phones in filled pauses (such as "um" and "uh") from those elsewhere, since it has been shown in previous work that durations of such fillers (which are very frequent in Switchboard) are considerably longer than those of spectrally similar vowels elsewhere (Shriberg, 1999). We also noted that for some phones, particularly nasals, errors in the recognizer forced alignments in training sometimes produced inordinately long (incorrect) phone durations. This affected the robustness of our standard deviation estimates; to avoid the problem we removed any clear outliers by inspecting the phone-specific duration histograms prior to computing standard deviations.

In addition to using phone-specific means and standard deviations over all speakers in a corpus,

---

[1] Improvements in future work could include the use of triphone-based normalization (on a sufficiently large corpus to assure robust estimates), or of normalization based on syllable position and stress information (given a dictionary marked for this information).

[2] Using dictionary stress information would probably be a better approach. Nevertheless, one advantage of this simple method is a robustness to pronunciation variation, since the longest observed normalized phone duration is used, rather than some predetermined phone.

we investigated the use of speaker-specific values for normalization, backing off to cross-speaker values for cases of low phone-by-speaker counts. However, these features were less useful than the features from data pooled over all speakers (probably due to a lack of robustness in estimating the standard deviations in the smaller, speaker-specific data sets). Alternative normalizations were also computed, including $phone\_dur_i/mean\_phone\_dur_i$ (to avoid noisy estimates of standard deviations), both for speaker-independent and speaker-dependent means.

Interestingly, we found it necessary to bin the normalized duration measures in order to reflect pre-boundary lengthening, rather than segmental information. Because these duration measures were normalized by phone-specific values (means and standard deviations), our decision trees were able to use certain specific feature values as clues to word identities and, indirectly, to boundaries. For example, the word "I" in the Switchboard corpus is a strong cue to a sentence onset; normalizing by the constant mean and standard deviation for that particular vowel resulted in specific values that were "learned" by the models. To address this, we binned all duration features to remove the level of precision associated with the phone-level correlations.

2.1.2.3   F0 features.   Pitch information is typically less robust and more difficult to model than other prosodic features, such as duration. This is largely attributable to variability in the way pitch is used across speakers and speaking contexts, complexity in representing pitch patterns, segmental effects, and pitch tracking discontinuities (such as doubling errors and pitch halving, the latter of which is also associated with nonmodal voicing).

To smooth out microintonation and tracking errors, simplify our F0 feature computation, and identify speaking-range parameters for each speaker, we post-processed the frame-level F0 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). Postprocessing steps are outlined in Fig. 2 and are described further in work on prosodic modeling for speaker verification (Sönmez et al., 1998).

The raw pitch tracker output has two main noise sources, which are minimized in the filtering stage. F0 halving and doubling are estimated by a lognormal tied mixture model (LTM) of F0, based on histograms of F0 values collected from all data from the same speaker.[3] For the Broadcast News corpus we pooled data from the same speaker over multiple news shows; for the Switchboard data, we used only the data from one side of a conversation for each histogram.

For each speaker, the F0 distribution was modeled by three lognormal modes spaced $\log 2$ apart in the log frequency domain. The locations of the modes were modeled with one tied parameter $(\mu - \log 2, \mu, \mu + \log 2)$, variances were scaled to be the same in the log domain, and mixture weights were estimated by an expectation maximization (EM) algorithm. This approach allowed estimation of speaker F0 range parameters that proved useful for F0 normalization.

Prior to the regularization stage, median filtering smooths voicing onsets during which the tracker is unstable, resulting in local undershoot or overshoot. We applied median filtering to windows of voiced frames with a neighborhood size of 7 plus or minus 3 frames. Next, in the regularization stage, F0 contours are fit by a simple piecewise linear model

$$\tilde{F}_0 = \sum_{k=1}^{K}(a_k F_0 + b_k)I_{[x_{k-1} < F_0 \le x_k]}$$

where K is the number of nodes, x_k are the node locations, and a_k and b_k are the linear parameters for a given region. The parameters are estimated by minimizing the mean squared error with a greedy node placement algorithm. The smoothness of the fits is fixed by two global parameters: the maximum mean squared error for deviation from a line in a given region, and the minimum length of a region.

The resulting filtered and stylized F0 contour, an example of which is shown in Fig. 3, enables robust extraction of features such as the value of the F0 slope at a particular point, the maximum or minimum stylized F0 within a region, and a simple characterization of whether the F0 trajectory before a word boundary is broken or continued into the next word. In addition, over all data from a particular speaker, statistics

---

[3]We settled on a cheating approach here, assuming speaker tracking information was available in testing, since automatic speaker segmentation and tracking was beyond the scope of this work.
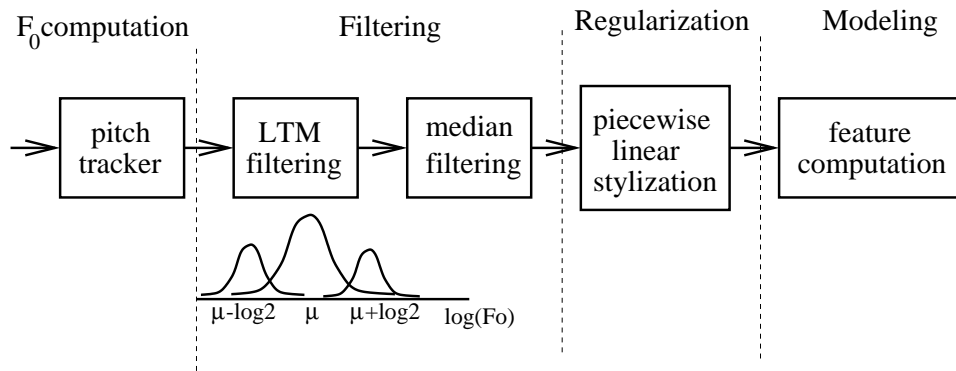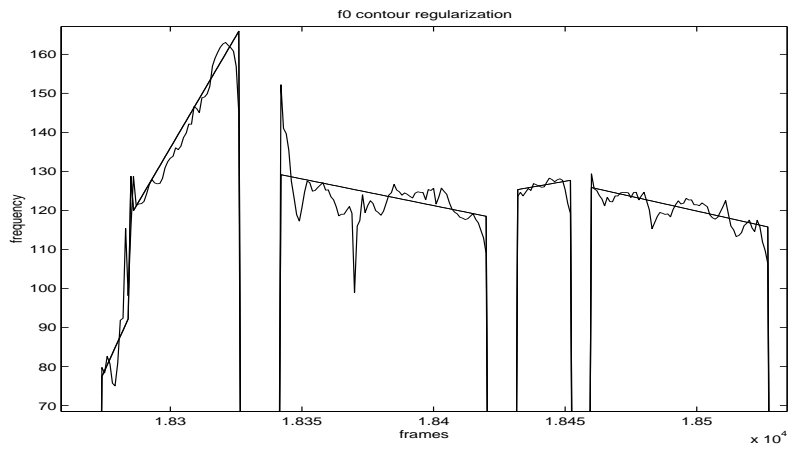
Fig. 2: F0 processing

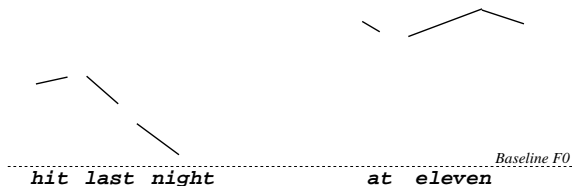Fig. 3: F0 contour filtering and regularization

9

Baseline F0

hit last night   at eleven

Fig. 4: Schematic example of stylized F0 for voiced regions of the text. The speaker's estimated baseline F0 (from the lognormal tied mixture modeling) is also indicated.

such as average slopes can be computed for normalization purposes. These statistics, combined with the speaker range values computed from the speaker histograms, allowed us to easily and robustly compute a large number of F0 features, as outlined in Section 2.1.2. In exploratory work on Switchboard, we found that the stylized F0 features yielded better results than more complex features computed from the raw F0 tracks. Thus, we restricted our input features to those computed from the processed F0 tracks, and did the same for Broadcast News.

We computed four different types of F0 features, all based on values computed from the stylized processing, but each capturing a different aspect of intonational behavior: (1) F0 *reset* features, (2) F0 *range* features, (3) F0 *slope* features, and (4) F0 *continuity* features. The general characteristics captured can be illustrated with the help of Fig. 4.

Reset features. The first set of features was designed to capture the well-known tendency of speakers to reset pitch at the start of a new major unit, such as a topic or sentence boundary, relative to where they left off. Typically the reset is preceded by a final fall in pitch associated with the ends of such units. Thus, at boundaries we expect a larger reset than at nonboundaries. We took measurements from the stylized F0 contours for the voiced regions of the word preceding and of the word following the boundary. Measurements were taken at either the minimum, maximum, mean, starting, or ending stylized F0 value within the region associated with each of the words. Numerous features were computed to compare the previous

to the following word; we computed both the log of the ratio between the two values, and the log of the difference between them, since it is unclear which measure would be better. Thus, in Fig. 4, the F0 difference between "at" and "eleven" would not imply a reset, but that between "night" and "at" would imply a large reset, particularly for the measure comparing the minimum F0 of "night" to the maximum F0 of "at". Parallel features were also computed based on the 200 ms windows rather than the words.

Range features. The second set of features reflected the pitch range of a single word (or window), relative to one of the speaker-specific global F0 range parameters computed from the lognormal tied mixture modeling described earlier. We looked both before and after the boundary, but found features of the pre-boundary word or window to be the most useful for these tasks. For the speaker-specific range parameters, we estimated F0 baselines, toplines, and some intermediate range measures. By far the most useful value in our modeling was the F0 baseline, which we computed as occurring halfway between the first mode and the second mode in each speaker-specific F0 histogram, i.e., roughly at the bottom of the modal (nonhalved) speaking range. We also estimated F0 toplines and intermediate values in the range, but these parameters proved much less useful than the baselines across tasks.

Unlike the reset features, which had to be defined as "missing" at boundaries containing a speaker change, the range features are defined at all boundaries for which F0 estimates can be made (since they look only at one side of the boundary). Thus for example in Fig. 4, the F0 of the word "night" falls very close to the speaker's F0 baseline, and can be utilized irrespective of whether or not the speaker changes before the next word.

We were particularly interested in these features for the case of topic segmentation in Broadcast News, since due to the frequent speaker changes at actual topic boundaries we needed a measure that would be defined at such locations. We also expected speakers to be more likely to fall closer to the bottom of their pitch range for topic than for sentence boundaries, since the former implies a greater degree of finality.

Slope features. Our final two sets of F0 features looked at the slopes of the stylized F0 segments, both for a word (or window) on only one side of the bound-

ary, and for continuity across the boundary. The aim was to capture local pitch variation such as the presence of pitch accents and boundary tones. Slope features measured the degree of F0 excursion before or after the boundary (relative to the particular speaker's average excursion in the pitch range), or simply normalized by the pitch range on the particular word.

Continuity features. Continuity features measured the change in slope across the boundary. Here, we expected that continuous trajectories would correlate with nonboundaries, and broken trajectories would tend to indicate boundaries, regardless of difference in pitch values across words. For example, in Fig. 4 the words "last" and "night" show a continuous pitch trajectory, so that it is highly unlikely there is a major syntactic or semantic boundary at that location. We computed both scalar (slope difference) and categorical (rise-fall) features for inclusion in the experiments.

2.1.2.4 Estimated voice quality features. Scalar F0 statistics (e.g., those contributing to slopes, or minimum/maximum F0 within a word or region) were computed ignoring any frames associated with F0 halving or doubling (frames whose highest posterior was not that for the modal region). However, regions corresponding to F0 halving as estimated by the lognormal tied mixture model showed high correlation with regions of creaky voice or glottalization that had been independently hand-labeled by a phonetician. Since creak may correlate with our boundaries of interest, we also included some categorical features, reflecting the presence or absence of creak.

We used two simple categorical features. One feature reflected whether or not pitch halving (as estimated by the model) was present for at least a few frames, anywhere within the word preceding the boundary. The second version looked at whether halving was present at the end of that word. As it turned out, while these two features showed up in decision trees for some speakers, and in the patterns we expected, glottalization and creak are highly speaker dependent and thus were not helpful in our overall modeling. However, for speaker-dependent modeling, such features could potentially be more useful.

2.1.2.5 Other features. We included two types of nonprosodic features, turn-related features and gender features. Both kinds of features were legitimately

available for our modeling, in the sense that standard speech recognition evaluations made this information known. Whether or not speaker change markers would actually be available depends on the application. It is not unreasonable however to assume this information, since automatic algorithms have been developed for this purpose (e.g., Przybocki and Martin, 1999; Liu and Kubala, 1999; Sönmez et al., 1999). Such nonprosodic features often interact with prosodic features. For example, turn boundaries cause certain prosodic features (such as F0 difference across the boundary) to be undefined, and speaker gender is highly correlated with F0. Thus, by including the features we could better understand feature interactions and check for appropriateness of normalization schemes.

Our turn-related features included whether or not the speaker changed at a boundary, the time elapsed from the start of the turn, and the turn count in the conversation. The last measure was included to capture structure information about the data, such as the preponderance of topic changes occurring early in Broadcast News shows, due to short initial summaries of topics at the beginning of certain shows.

We included speaker gender mainly as a check to make sure the F0 processing was normalized properly for gender differences. That is, we initially hoped that this feature would *not* show up in the trees. However, we learned that there are reasons other than poor normalization for gender to occur in the trees, including potential truly stylistic differences between men and women, and structure differences associated with gender (such as differences in lengths of stories in Broadcast News). Thus, gender revealed some interesting inherent interactions in our data, which are discussed further in Section 3.3. In addition to speaker gender, we included the gender of the listener, to investigate the degree to which features distinguishing boundaries might be affected by sociolinguistic variables.

2.1.3 Decision trees

As in past prosodic modeling work (Shriberg et al., 1997), we chose to use CART-style decision trees (Breiman et al., 1984), as implemented by the IND package (Buntine and Caruana, 1992). The software offers options for handling missing feature values (important since we did not have good pitch estimates for all data points), and is capable of processing large amounts of training data. Decision trees are prob-

11

abilistic classifiers that can be characterized briefly as follows. Given a set of discrete or continuous features and a labeled training set, the decision tree construction algorithm repeatedly selects a single feature that, according to an information-theoretic criterion (entropy), has the highest predictive value for the classification task in question.[4] The feature queries are arranged in a hierarchical fashion, yielding a tree of questions to be asked of a given data point. The leaves of the tree store probabilities about the class distribution of all samples falling into the corresponding region of the feature space, which then serve as predictors for unseen test samples. Various smoothing and pruning techniques are commonly employed to avoid overfitting the model to the training data.

Although any of several probabilistic classifiers (such as neural networks, exponential models, or naive Bayes networks) could be used as posterior probability estimators, decision trees allow us to add, and automatically select, other (nonprosodic) features that might be relevant to the task—including categorical features. Furthermore, decision trees make no assumptions about the shape of feature distributions; thus it is not necessary to convert feature values to some standard scale. And perhaps most importantly, decision trees offer the distinct advantage of interpretability. We have found that human inspection of feature interactions in a decision tree fosters an intuitive understanding of feature behaviors and the phenomena they reflect. This understanding is crucial for progress in developing better features, as well as for debugging the feature extraction process itself.

The decision tree served as a prosodic model for estimating the posterior probability of a (sentence or topic) boundary at a given inter-word boundary, based on the automatically extracted prosodic features. We define F_i as the features extracted from a window around the ith potential boundary, and T_i as the boundary type (boundary/no-boundary) at that position. For each task, decision trees were trained to predict the ith boundary type, i.e., to estimate $P(T_i|F_i, W)$. By design, this decision was only weakly conditioned on the word sequence W, insofar as some of the prosodic features depend on the phonetic alignment of the word models. We preferred the weak conditioning for ro-

bustness to word errors in speech recognizer output. Missing feature values in F_i occurred mainly for the F0 features (due to lack of robust pitch estimates for an example), but also at locations where features were inherently undefined (e.g., pauses at turn boundaries). Such cases were handled in testing by sending the test sample down each tree branch with the proportion found in the training set at that node, and then averaging the corresponding predictions.

2.1.4 Feature selection algorithm

Our initial feature sets contained a high degree of feature redundancy because, for example, similar features arose from changing only normalization schemes, and others (such as energy and F0) are inherently correlated in speech production. The greedy nature of the decision tree learning algorithm implies that larger initial feature sets can yield suboptimal results. The availability of more features provides greater opportunity for "greedy" features to be chosen; such features minimize entropy locally but are suboptimal with respect to entropy minimization over the whole tree. Furthermore, it is desirable to remove redundant features for computational efficiency and to simplify interpretation of results.

To automatically reduce our large initial candidate feature set to an optimal subset, we developed an iterative feature selection algorithm that involved running multiple decision trees in training (sometimes hundreds for each task). The algorithm combines elements of brute-force search with previously determined human-based heuristics for narrowing the feature space to good groupings of features. We used the entropy reduction of the overall tree after cross-validation as a criterion for selecting the best subtree. Entropy reduction is the difference in test-set entropy between the prior class distribution and the posterior distribution estimated by the tree. It is a more fine-grained metric than classification accuracy, and is thus the more appropriate measure to use for any of the model combination approaches described in Section 2.3.

The algorithm proceeds in two phases. In the first phase, the large number of initial candidate features is reduced by a leave-one-out procedure. Features that do not reduce performance when removed are eliminated from further consideration. The second phase begins with the reduced number of features,

---

[4]For multivalued or continuous features, the algorithm also determines optimal feature value subsets or thresholds, respectively, to compare the feature to.

and performs a beam search over all possible subsets of features. Because our initial feature set contained over 100 features, we split the set into smaller subsets based on our experience with feature behaviors. For each subset we included a set of "core" features, which we knew from human analyses of results served as catalysts for other features. For example, in all subsets, pause duration was included, since without this feature present, duration and pitch features are much less discriminative for the boundaries of interest.[5]

2.2 Language modeling

The goal of language modeling for our segmentation tasks is to capture information about segment boundaries contained in the word sequences. We denote boundary classifications by $T = T_1, \ldots, T_K$ and use $W = W_1, \ldots, W_N$ for the word sequence. Our general approach is to model the joint distribution of boundary types and words in a hidden Markov model (HMM), the hidden variable in this case being the boundaries T_i (or some related variable from which T_i can be inferred). Because we had hand-labeled training data available for all tasks, the HMM parameters could be trained in supervised fashion.

The structure of the HMM is task specific, as described below, but in all cases the Markovian character of the model allows us to efficiently perform the probabilistic inferences desired. For example, for topic segmentation we extract the most likely *overall* boundary classification

$$\operatorname*{argmax}_{T} P(T|W) \quad, \tag{2}$$

using the Viterbi algorithm (Viterbi, 1967). This optimization criterion is appropriate because the topic segmentation evaluation metric prescribed by the TDT program (Doddington, 1998) rewards overall consistency of the segmentation.[6]

For sentence segmentation, the evaluation metric simply counts the number of correctly labeled boundaries (see Section 2.4.4). Therefore, it is advantageous

---

[5]The success of this approach depends on the makeup of the initial feature sets, since highly correlated useful features can cancel each other out during the first phase. This problem can be addressed by forming initial feature subsets that minimize within-set cross-feature correlations.

[6]For example, given three sentences $s_1 s_2 s_3$ and strong evidence that there is a topic boundary between s_1 and s_3, it is better to output a boundary either before or after s_2, but not in both places.

to use the slightly more complex forward-backward algorithm (Baum et al., 1970) to maximize the posterior probability of each individual boundary classification T_i

$$\operatorname*{argmax}_{T_i} P(T_i|W) \quad. \tag{3}$$

This approach minimizes the expected per-boundary classification error rate (Dermatas and Kokkinakis, 1995).

2.2.1 Sentence segmentation

We relied on a hidden-event N-gram language model (LM) (Stolcke and Shriberg, 1996; Stolcke et al., 1998). The states of the HMM consist of the end-of-sentence status of each word (boundary or no-boundary), plus any preceding words and possibly boundary tags to fill up the N-gram context ($N = 4$ in our experiments). Transition probabilities are given by N-gram probabilities estimated from annotated, boundary-tagged training data using Katz backoff (Katz, 1987). For example, the bigram parameter $P(\texttt{<s>}|\texttt{tonight})$ gives the probability of a sentence boundary following the word "tonight". HMM observations consist of only the current word portion of the underlying N-gram state (with emission likelihood 1), constraining the state sequence to be consistent with the observed word sequence.

2.2.2 Topic segmentation

We first constructed 100 individual unigram topic cluster language models, using the multipass *k*-means algorithm described in (Yamron et al., 1998). We used the pooled Topic Detection and Tracking (TDT) Pilot and TDT-2 training data (Cieri et al., 1999). We removed stories with fewer than 300 and more than 3000 words, leaving 19,916 stories with an average length of 538 words. Then, similar to the Dragon topic segmentation approach (Yamron et al., 1998), we built an HMM in which the states are topic clusters, and the observations are sentences. The resulting HMM forms a complete graph, allowing transition between any two topic clusters. In addition to the basic HMM segmenter, we incorporated two states for modeling the initial and final sentences of a topic segment. We reasoned that this can capture formulaic speech patterns used by broadcast speakers. Likelihoods for the

start and end models are obtained as the unigram language model probabilities of the topic-initial and final sentences, respectively, in the training data. Note that single start and end states are shared for all topics, and traversal of the initial and final states is optional in the HMM topology. The topic cluster models work best if whole blocks of words or "pseudo-sentences" are evaluated against the topic language models (the likelihoods are otherwise too noisy). We therefore presegment the data stream at pauses exceeding 0.65 second, as process we will refer to as "chopping".

2.3 Model combination

We expect prosodic and lexical segmentation cues to be partly complementary, so that combining both knowledge sources should give superior accuracy over using each source alone. This raises the issue of how the knowledge sources should be integrated. Here, we describe two approaches to model combination that allow the component prosodic and lexical models to be retained without much modification. While this is convenient and computationally efficient, it prevents us from explicitly modeling interactions (i.e., statistical dependence) between the two knowledge sources. Other researchers have proposed model architectures based on decision trees (Heeman and Allen, 1997) or exponential models (Beeferman et al., 1999) that can potentially integrate the prosodic and lexical cues discussed here. In other work (Stolcke et al., 1998; Tür et al., 2000) we have started to study integrated approaches for the segmentation tasks studied here, although preliminary results show that the simple combination techniques are very competitive in practice.

2.3.1 Posterior probability interpolation

Both the prosodic decision tree and the language model (via the forward-backward algorithm) estimate posterior probabilities for each boundary type T_i. We can arrive at a better posterior estimator by linear interpolation:

$$P(T_i|W,F) \approx \lambda P_{\text{LM}}(T_i|W) + (1-\lambda)P_{\text{DT}}(T_i|F_i,W) \tag{4}$$

where λ is a parameter optimized on held-out data to optimize the overall model performance.

2.3.2 Integrated hidden Markov modeling

Our second model combination approach is based on the idea that the HMM used for lexical modeling can be extended to "emit" both words and prosodic observations. The goal is to obtain an HMM that models the joint distribution $P(W,F,T)$ of word sequences W, prosodic features F, and hidden boundary types T in a Markov model. With suitable independence assumptions we can then apply the familiar HMM techniques to compute

$$\operatorname*{argmax}_{T} P(T|W,F)$$

or

$$\operatorname*{argmax}_{T_i} P(T_i|W,F) \quad,$$

which are now conditioned on both lexical and prosodic cues. We describe this approach for sentence segmentation HMMs; the treatment for topic segmentation HMMs is mostly analogous but somewhat more involved, and described in detail elsewhere (Tür et al., 2000).

To incorporate the prosodic information into the HMM, we model prosodic features as emissions from relevant HMM states, with likelihoods $P(F_i|T_i,W)$, where F_i is the feature vector pertaining to potential boundary T_i. For example, an HMM state representing a sentence boundary <S> at the current position would be penalized with the likelihood $P(F_i|\texttt{<S>})$. We do so based on the assumption that prosodic observations are conditionally independent of each other given the boundary types T_i and the words W. Under these assumptions, a complete path through the HMM is associated with the total probability

$$P(W,T)\prod_i P(F_i|T_i,W) = P(W,F,T) \quad, \tag{5}$$

as desired.

The remaining problem is to estimate the likelihoods $P(F_i|T_i,W)$. Note that the decision tree estimates posteriors $P_{\text{DT}}(T_i|F_i,W)$. These can be converted to likelihoods using Bayes' rule as in

$$P(F_i|T_i,W) = \frac{P(F_i|W)P_{\text{DT}}(T_i|F_i,W)}{P(T_i|W)} \quad. \tag{6}$$

The term $P(F_i|W)$ is a constant for all choices of T_i and can thus be ignored when choosing the most probable one. Next, because our prosodic model is purposely not conditioned on word identities, but only on

aspects of W that relate to time alignment, we approximate $P(T_i|W) \approx P(T_i)$. Instead of explicitly dividing the posteriors, we prefer to downsample the training set to make $P(T_i = \mathbf{yes}) = P(T_i = \mathbf{no}) = \frac{1}{2}$. A beneficial side effect of this approach is that the decision tree models the lower-frequency events (segment boundaries) in greater detail than if presented with the raw, highly skewed class distribution.

When combining probabilistic models of different types, it is advantageous to weight the contributions of the language models and the prosodic trees relative to each other. We do so by introducing a tunable *model combination weight* (MCW), and by using $P_{\mathrm{DT}}(F_i|T_i, W)^{\mathrm{MCW}}$ as the effective prosodic likelihoods. The value of MCW is optimized on held-out data.

2.3.3   HMM posteriors as decision tree features

A third approach could be used to combine the language and prosodic models, although for practical reasons we chose not to use it in this work. In this approach, an HMM incorporating only lexical information is used to compute posterior probabilties of boundary types, as described in Section 2.3.1. A prosodic decision tree is then trained, using the HMM posteriors as additional input features. The tree is free to combine the word-based posteriors with prosodic features; it can thus model limited forms of dependence between prosodic and word-based information (as summarized in the posteriors).

A severe drawback of using posteriors in the decision tree, however, is that in our current paradigm, the HMM is trained on correct words. In testing, the tree may therefore grossly overestimate the informativeness of the word-based posteriors based on automatic transcriptions. Indeed, we found that on a hidden-event detection task similar to sentence segmentation (Stolcke et al., 1998) this model combination method worked well on true words, but faired worse than the other approaches on recognized words. To remedy the mismatch between training and testing of the combined model, we would have to train, as well as test, on recognized words; this would require computationally intensive processing of a large corpus. For these reasons, we decided not to use HMM posteriors as tree features in the present studies.

2.3.4   Alternative models

A few additional comments are in order regarding our choice of model architectures and possible alternatives. The HMMs used for lexical modeling are *likelihood models*, i.e., they model the probabilities of observations given the hidden variables (boundary types) to be inferred, while making assumptions about the independence of the observations given the hidden events. The main virtue of HMMs in our context is that they integrate the local evidence (words and prosodic features) with models of context (the N-gram history) in a very computationally efficient way (for both training and testing). A drawback is that the independence assumptions may be inappropriate and may therefore inherently limit the performance of the model.

The decision trees used for prosodic modeling, on the other hand, are *posterior models*, i.e., they directly model the probabilities of the unknown variables given the observations. Unlike likelihood-based models, this has the advantages that model training explicitly enhances discrimination between the target classifications (i.e., boundary types), and that input features can be combined easily to model interactions between them. Drawbacks are the sensitivity to skewed class distributions (as pointed out in the previous section), and the fact that it becomes computationally expensive to model interactions between multiple target variables (e.g., adjacent boundaries). Furthermore, input features with large discrete ranges (such as the set of words) present practical problems for many posterior model architectures.

Even for the tasks discussed here, other modeling choices would have been practical, and await comparative study in future work. For example, posterior lexical models (such as decision trees or neural network classifiers) could be used to predict the boundary types from words and prosodic features together, using word-coding techniques developed for tree-based language models (Bahl et al., 1989). Conversely, we could have used prosodic likelihood models, removing the need to convert posteriors to likelihoods. For example, the continuous feature distributions could be modeled with (mixtures of) multidimensional Gaussians (or other types of distributions), as is commonly done for the spectral features in speech recognizers (Digalakis and Murveit, 1994, among others).

2.4   Data

2.4.1   Speech data and annotations

Switchboard data used in sentence segmentation was drawn from a subset of the corpus (Godfrey et al., 1992) that had been hand-labeled for sentence boundaries (Meteer et al., 1995) by the Linguistic Data Consortium (LDC). Broadcast News data for topic and sentence segmentation was extracted from the LDC's 1997 Broadcast News (BN) release. Sentence boundaries in BN were automatically determined using the MITRE sentence tagger (Palmer and Hearst, 1997) based on capitalization and punctuation in the transcripts. Topic boundaries were derived from the SGML markup of story units in the transcripts. Training of Broadcast News language models for sentence segmentation also used an additional 130 million words of text-only transcripts from the 1996 Hub-4 language model corpus, in which sentence boundaries had been marked by SGML tags.

2.4.2   Training, tuning, and test sets

Table 1 shows the amount of data used for the various tasks. For each task, separate datasets were used for model training, for tuning any free parameters (such as the model combination and posterior interpolation weights), and for final testing. In most cases the language model and the prosodic model components used different amounts of training data.

As is common for speech recognition evaluations on Broadcast News, frequent speakers (such as news anchors) appear in both training and test sets. By contrast, in Switchboard our train and test sets did not share any speakers. In both corpora, the average word count per speaker decreased roughly monotonically with the percentage of speakers included. In particular, the Broadcast News data contained a large number of speakers who contributed very few words. A reasonably meaningful statistic to report for words per speaker is thus a weighted average, or the average number of datapoints by the same speaker. On that measure, the two corpora had similar statistics: 6687.11 and 7525.67 for Broadcast News and Switchboard, respectively.

2.4.3   Word recognition

Experiments involving recognized words used the 1-best output from SRI's DECIPHER large-vocabulary speech recognizer. We simplified processing by skipping several of the computationally expensive or cumbersome steps often used for optimum performance, such as acoustic adaptation and multiple-pass decoding. The recognizer performed one bigram decoding pass, followed by a single N-best rescoring pass using a higher-order language model. The Switchboard test set was decoded with a word error rate of 46.7% using acoustic models developed for the 1997 Hub-5 evaluation (National Institute for Standards and Technology, 1997). The Broadcast News recognizer was based on the 1997 SRI Hub-4 recognizer (Sankar et al., 1998) and had a word error rate of 30.5% on the test set used in our study.

2.4.4   Evaluation metrics

Sentence segmentation performance for true words was measured by boundary classification error, i.e. the percentage of word boundaries labeled with the incorrect class. For recognized words, we first performed a string alignment of the automatically labeled recognition hypothesis with the reference word string (and its segmentation). Based on this alignment we then counted the number of incorrectly labeled, deleted, and inserted word boundaries, expressed as a percentage of the total number of word boundaries. This metric yields the same result as the boundary classification error rate if the word hypothesis is correct. Otherwise, it includes additional errors from inserted or deleted boundaries, in a manner similar to standard word error scoring in speech recognition. Topic segmentation was evaluated using the metric defined by NIST for the TDT-2 evaluation (Doddington, 1998).

3   Results and discussion

The following sections describe results from the prosodic modeling approach, for each of our three tasks. The first three sections focus on the tasks individually, detailing the features used in the best-performing tree. For sentence segmentation, we report on trees trained on non-downsampled data, as used in the posterior interpolation approach. For all tasks,

16

Table 1: Size of speech data sets used for model training and testing for the three segmentation tasks

Task	Training		Tuning	Test
	LM	Prosody		
SWB Sentence (transcribed)	1788 sides (1.2M words)	1788 sides (1.2M words)	209 sides (103K words)	209 sides (101K words)
SWB Sentence (recognized)	1788 sides (1.2M words)	1788 sides (1.2M words)	12 sides (6K words)	38 sides (18K words)
BN Sentence	103 shows + BN96 (130M words)	93 shows (700K words)	5 shows (24K words)	5 shows (21K words)
BN Topic	TDT + TDT2 (10.7M words)	93 shows (700K words)	10 shows (205K words)	6 shows (44K words)

including topic segmentation, we also trained down-sampled trees for the HMM combination approach. Where both types of trees were used (sentence segmentation), feature usage on downsampled trees was roughly similar to that of the non-downsampled trees, so we describe only the non-downsampled trees. For topic segmentation, the description refers to a downsampled tree.

In each case we then look at results from combining the prosodic information with language model information, for both transcribed and recognized words. Where possible (i.e., in the sentence segmentation tasks), we compare results for the two alternative model integration approaches (combined HMM and interpolation). In the next two sections, we compare results across both tasks and speech corpora. We discuss differences in which types of features are helpful for a task, as well as differences in the relative reduction in error achieved by the different models, using a measure that tries to normalize for the inherent difficulty of each task. Finally, we discuss issues for future work.

3.1 Task 1: Sentence segmentation of Broadcast News data

3.1.1 Prosodic feature usage

The best-performing tree identified six features for this task, which fall into four groups. To summarize the relative importance of the features in the decision tree we use a measure we call "feature usage", which is computed as the relative frequency with which that feature or feature class is queried in the decision tree.

The measure increments for each sample classified using that feature; features used higher in the tree classify more samples and therefore have higher usage values. The feature usage was as follows (by type of feature):

- (46%) Pause duration at boundary

- (42%) Turn/no turn at boundary

- (11%) F0 difference across boundary

- (01%) Rhyme duration

The main features queried were pause, turn, and F0. To understand whether they behaved in the manner expected based on the descriptive literature, we inspected the decision tree. The tree for this task had 29 leaves; we show the top portion of it in Fig. 5.

The behavior of the features is precisely that expected from the literature. Longer pause durations at the boundary imply a higher probability of a sentence boundary at that location. Speakers exchange turns almost exclusively at sentence boundaries in this corpus, so the presence of a turn boundary implies a sentence boundary. The F0 features all behave in the same way, with lower negative values raising the probability of a sentence boundary. These features reflect the log of the ratio of F0 measured within the word (or window) preceding the boundary to the F0 in the word (or window) after the boundary. Thus, lower negative values imply a larger pitch reset at the boundary, consistent with what we would expect.
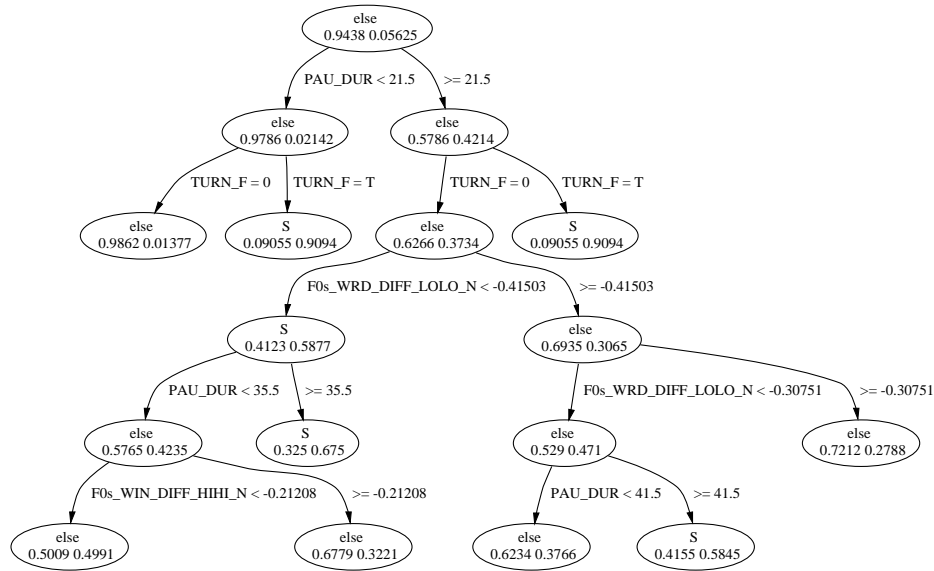
17

else
0.9438 0.05625

PAU_DUR < 21.5   >= 21.5

else
0.9786 0.02142

else
0.5786 0.4214

TURN_F = 0   TURN_F = T

TURN_F = 0   TURN_F = T

else
0.9862 0.01377

S
0.09055 0.9094

else
0.6266 0.3734

S
0.09055 0.9094

F0s_WRD_DIFF_LOLO_N < -0.41503   >= -0.41503

S
0.4123 0.5877

else
0.6935 0.3065

PAU_DUR < 35.5   >= 35.5

F0s_WRD_DIFF_LOLO_N < -0.30751   >= -0.30751

else
0.5765 0.4235

S
0.325 0.675

else
0.529 0.471

else
0.7212 0.2788

F0s_WIN_DIFF_HIHI_N < -0.21208   >= -0.21208

PAU_DUR < 41.5   >= 41.5

else
0.5009 0.4991

else
0.6779 0.3221

else
0.6234 0.3766

S
0.4155 0.5845

Fig. 5: Top levels of decision tree selected for the Broadcast News sentence segmentation task. Nodes contain the percentage of "else" and "S" (sentence) boundaries, respectively, and are labeled with the majority class. PAU_DUR=pause duration, F0s=stylized F0 feature reflecting ratio of speech before the boundary to that after that boundary, in the log domain.

3.1.2 Error reduction from prosody

Table 2 summarizes the results on both transcribed and recognized words, for various sentence segmentation models for this corpus. The baseline (or "chance") performance for true words in this task is 6.2% error, obtained by labeling all locations as nonboundaries (the most frequent class). For recognized words, it is considerably higher; this is due to the non-zero lower bound resulting if one accounts for locations in which the 1-best hypothesis boundaries do not coincide with those of the reference alignment. "Lower bound" gives the lowest segmentation error rate possible given the word boundary mismatches due to recognition errors.

Results show that the prosodic model alone performs better than a word-based language model, despite the fact that the language model was trained on a much larger data set. Furthermore, the prosodic model is somewhat more robust to errorful recognizer output than the language model, as measured by the absolute increase in error rate in each case. Most importantly, a statistically significant error reduction is achieved by combining the prosodic features with the lexical features, for both integration methods. The relative error reduction is 19% for true words, and 8.5% for recognized words. This is true even though both models contained turn information, thus violating the independence assumption made in the model combination.

3.1.3 Performance without F0 features

A question one may ask in using the prosody features, is how the model would perform without any F0 features. Unlike pause, turn, and duration information, the F0 features used are not typically extracted or computed in most ASR systems. We ran comparison experiments on all conditions, but removing all F0 features from the input to the feature selection algorithm. Results are shown in Table 3, along with the previous results using all features, for comparison.

As shown, the effect of removing F0 features reduces model accuracy for prosody alone, for both true and recognized words. In the case of the true words, model integration using the no-F0 prosodic tree actually fares slightly better than that which used all features, despite similar model combination weights in the two cases. The effect is only marginally signifi-

cant in a Sign test, so it may indicate chance variation. However it could also indicate a higher degree of correlation between true words and the prosodic features that indicate boundaries, when F0 is included. However, for recognized words, the model with all prosodic features is superior to that without the F0 features, both alone and after integration with the language model.

3.2 Task 2: Sentence segmentation of Switchboard data

3.2.1 Prosodic feature usage

Switchboard sentence segmentation made use of a markedly different distribution of features than observed for Broadcast News. For Switchboard, the best-performing tree found by the feature selection algorithm had a feature usage as follows:

- (49%) Phone and rhyme duration preceding boundary

- (18%) Pause duration at boundary

- (17%) Turn/no turn at boundary

- (15%) Pause duration at *previous* word boundary

- (01%) Time elapsed in turn

Clearly, the primary feature type used here is pre-boundary duration, a measure that was used only a scant 1% of the time for the same task in news speech. Pause duration at the boundary was also useful, but not to the degree found for Broadcast News.

Of course, it should be noted in comparing feature usage across corpora and tasks that results here pertain to comparisons of *the most parsimonious, best-performing model* for each corpus and task. That is, we do not mean to imply that an individual feature such as preboundary duration is not useful in Broadcast News, but rather that the minimal and most successful model for that corpus makes little use of that feature (because it can make better use of other features). Thus, it cannot be inferred from these results that some feature not heavily used in the minimal model is not helpful. The feature may be useful on

Table 2: Results for sentence segmentation on Broadcast News

Model	Transcribed words	Recognized words
LM only (130M words)	4.1	11.8
Prosody only (700K words)	3.6	10.9
Interpolated	3.5	10.8
Combined HMM	3.3	11.7
Chance	6.2	13.3
Lower bound	0.0	7.9

Values are word boundary classification error rates (in percent).

Table 3: Results for sentence segmentation on Broadcast News, with and without F0 features

Model	Transcribed Words	Recognized Words
LM only (130M words)	4.1	11.8
All Prosody Features:		
Prosody only (700K words)	3.6	10.9
Prosody+LM: Combined HMM	3.3	
Prosody+LM: Interpolation		10.8
No F0 Features:		
Prosody only (700K words)	3.8	11.3
Prosody+LM: Combined HMM	3.2	
Prosody+LM: Interpolation		11.1
Chance	6.2	13.3
Lower bound	0.0	7.9

Values are word boundary classification error rates (in percent). For the integrated ("Prosody + LM") models, results are given for the optimal model only (combined HMM for true words, interpolation of posteriors for recognized words.)

its own; however, it is not as useful as some other feature(s) made available in this study.[7]

The two "pause" features are not grouped together, because they represent fundamentally different phenomena. The second pause feature essentially captured the boundaries after one word such as "uh-huh" and "yeah", which for this work had been marked as followed by sentence boundaries ("yeah `<Sent>` i know what you mean").[8] The previous pause in this case was time that the speaker had spent in listening to the other speaker (channels were recorded separately and recordings were continuous on both sides). Since one-word backchannels (acknowledgments such as "uh-huh") and other short dialogue acts make up a large percentage of sentence boundaries in this corpus, the feature is used fairly often. The turn features also capture similar phenomena related to turn-taking. The leaf count for this tree was 236, so we display only the top portion of the tree in Fig. 6.

Pause and turn information, as expected, suggested sentence boundaries. Most interesting about this tree was the consistent behavior of duration features, which gave higher probability to a sentence boundary when lengthening of phones or rhymes was detected in the word preceding the boundary. Although this is in line with descriptive studies of prosody, it was rather remarkable to us that duration would work at all, given the casual style and speaker variation in this corpus, as well as the somewhat noisy forced alignments for the prosodic model training.

3.2.2 Error reduction from prosody

Unlike the previous results for the same task on Broadcast News, we see in Table 4 that for Switchboard data, prosody alone is not a particularly good model. For transcribed words it is considerably worse than the language model; however, this difference is reduced for the case of recognized words (where the prosody shows less degradation than the language

Table 4: Results for sentence segmentation on Switchboard

Model	Transcribed words	Recognized words
LM only	4.3	22.8
Prosody only	6.7	22.9
Interpolated	4.1	22.2
Combined HMM	4.0	22.5
Chance	11.0	25.8
Lower bound	0.0	17.6

Values are word boundary classification error rates (in percent).

model).

Yet, despite the poor performance of prosody alone, combining prosody with the language model resulted in a statistically significant improvement over the language model alone (7.0% and 2.6% relative for true and recognized words, respectively). All differences were statistically significant, including the difference in performance between the two model integration approaches. Furthermore, the pattern of results for model combination approaches observed for Broadcast News holds as well: the combined HMM is superior for the case of transcribed words, but suffers more than the interpolation approach when applied to recognized words.

3.3 Task 3: Topic segmentation of Broadcast News data

3.3.1 Prosodic feature usage

The feature selection algorithm determined five feature types most helpful for this task:

- (43%) Pause duration at boundary
- (36%) F0 range
- (09%) Turn/no turn at boundary
- (07%) Speaker gender
- (05%) Time elapsed in turn

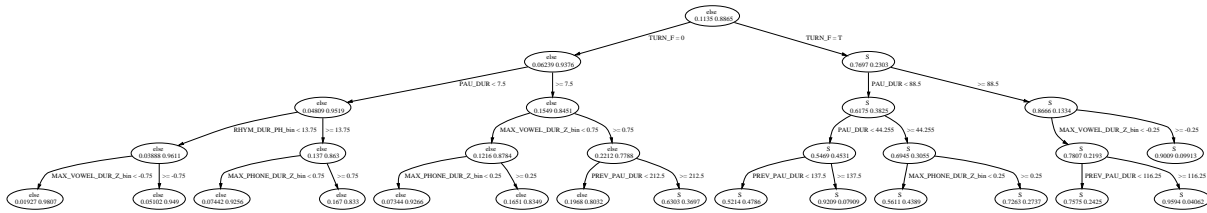The results are somewhat similar to those seen earlier for sentence segmentation in Broadcast News, in

---

[7] One might propose a more thorough investigation by reporting performance for one feature at a time. However, we found in examining such results that typically our features required the presence of one or more additional features in order to be helpful. (For example, pitch features required the presence of the pause feature.) Given the large number of features used, the number of potential combinations becomes too large to report on fully here.

[8] "Utterance" boundary is probably a better term, but for consistency we use the term "sentence" boundary for these dialogue act boundaries as well.

Fig. 6: Top levels of decision tree selected for the Switchboard sentence segmentation task. Nodes contain the percentage of "S" (sentence) and "else" boundaries, respectively, and are labeled with the majority class. "PAU_DUR"=pause duration, "RHYM"=syllable rhyme. VOWEL, PHONE and RHYME features apply to the word before the boundary.

that pause, turn, and F0 information are the top features. However, the feature usage here differs considerably from that for the sentence segmentation task, in that here we see a much higher use of F0 information.

Furthermore, the most important F0 feature was a range feature (log ratio of the preceding word's F0 to the speaker's F0 baseline), which was used 2.5 times more often in the tree than the F0 feature based on difference across the boundary. The range feature does not require information about F0 on the other side of the boundary; thus, it could be applied regardless of whether there was a speaker change at that location. This was a much more important issue for topic segmentation than for sentence segmentation, since the percentage of speaker changes is higher in the former than in the latter.

It should be noted, however, that the importance of pause duration is underestimated. As explained earlier, pause duration was also used *prior* to tree building, in the chopping process. The decision tree was applied only to boundaries exceeding a certain duration. Since the duration threshold was found by optimizing for the TDT error criterion, which assigns greater weight to false alarms than to false rejections, the resulting pause threshold is quite high (over half a second). Separate experiments using boundaries below our chopping threshold show that trees distinguish much shorter pause durations for segmentation decisions, implying that prosody could potentially yield an even larger relative advantage for error metrics favoring a shorter chopping threshold.

Inspecting the tree in Fig. 7 (the tree has additional leaves; we show only the top of it), we find that

it is easily interpretable and consistent with prosodic descriptions of topic or paragraph boundaries. Boundaries are indicated by longer pauses and by turn information, as expected. Note that the pause thresholds are considerably higher than those used for the sentence tree. This is as expected, because of the larger units used here, and due to the prior chopping at long pause boundaries for this task.

Most of the rest of the tree uses F0 information, in two ways. The most useful F0 range feature, *F0s_LR_MEAN_KBASELN*, computes the log of the ratio of the mean F0 in the last word to the speaker's estimated F0 baseline. As shown, lower values favor topic boundaries, which is consistent with speakers dropping to the bottom of their pitch ranges at the ends of topic units. The other F0 feature reflects the height of the last word relative to a speaker's estimated F0 range; smaller values thus indicate that a speaker is closer to his or her F0 floor, and as would be predicted, imply topic boundaries.

The speaker-gender feature was used in the tree in a pattern that at first suggested to us a potential problem with our normalizations. It was repeatedly used immediately after conditioning on the F0 range feature *F0s_LR_MEAN_KBASELN*. However, inspection of the feature value distributions by gender and by boundary class suggested that this was not a problem with normalization, as shown in Fig. 8.

As indicated, there was no difference by gender in the distribution of F0 values for the feature in the case of boundaries not containing a topic change. After normalization, both men and women ended nontopic boundaries in similar regions above their baselines.
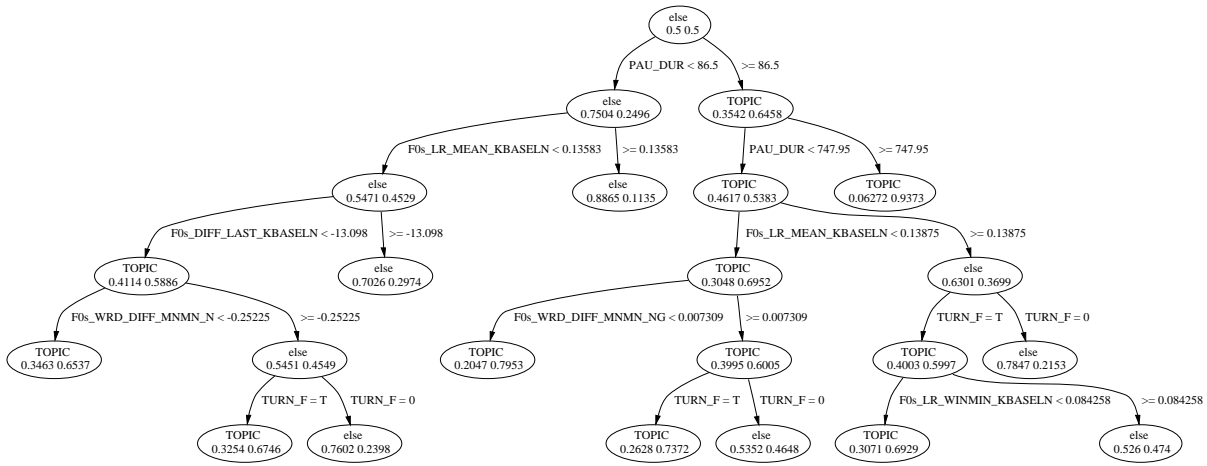
Fig. 7: Top levels of decision tree selected for the Broadcast News topic segmentation task. Nodes contain the percentage of "else" and "TOPIC" boundaries, respectively, and are labeled with the majority class.
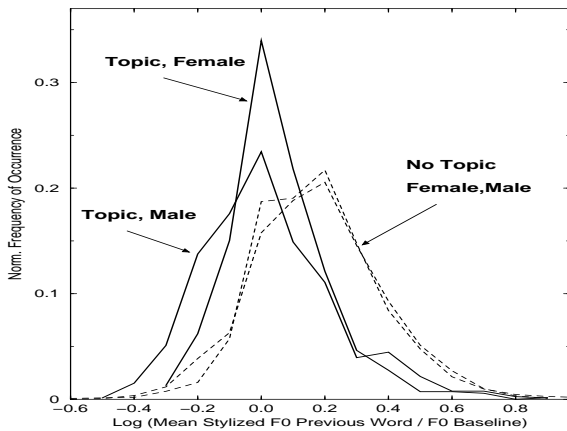

Fig. 8: Normalized distribution of F0 range feature (*F0s_LR_MEAN_KBASELN*) for male and female speakers for topic and nontopic boundaries in Broadcast News

Since nontopic boundaries are by far the more frequent class (distributions in the histogram are normalized), the majority of boundaries in the data show no difference on this measure by gender. For topic boundaries, however, the women in a sense behave more "neatly" than the men. As a group, the women have a tighter distribution, ending topics at F0 values that are centered closely around their F0 baselines. Men, on the other hand, are as a group somewhat less "well-behaved" in this regard. They often end topics below their F0 baselines, and showing a wider distribution (although it should also be noted that since these are aggregate distributions, the wider distribution for men could reflect either within-speaker or cross-speaker variation).

This difference is unlikely to be due to baseline estimation problems, since the nontopic distributions show no difference. The variance difference is also not explained by a difference in sample size, since that factor would predict an effect in the opposite direction. One possible explanation is that men are more likely than women to produce regions of nonmodal voicing (such as creak) at the ends of topic boundaries; this awaits further study. In addition, we noted that nontopic pauses (i.e., chopping boundaries) are much more likely to occur in male than in female speech, a phenomenon that could have several causes. For example, it could be that male speakers in Broadcast

23

Table 5: Results for topic segmentation on Broadcast News

Model	Transcribed words	Recognized words
LM only	0.1895	0.1897
Prosody only	0.1657	0.1731
Combined HMM	0.1377	0.1438
Chance	0.3	0.3

Values indicate the TDT weighted segmentation cost metric.

Table 6: Results for topic segmentation on Broadcast News

Model	Transcribed words
LM only	0.1895
Combined HMM:	
All prosodic features	0.1377
No F0 features	0.1511
Chance	0.3

Values indicate the TDT weighted segmentation cost metric.

News are assigned longer topic segments on average, or that male speakers are more prone to pausing in general, or that males dominate the spontaneous speech portions where pausing is naturally more frequent. This finding, too, awaits further analysis.

3.3.2 Error reduction from prosody

Table 5 shows results for segmentation into topics in Broadcast News speech. All results reflect the word-averaged, weighted error metric used in the TDT-2 evaluations (Doddington, 1998). Chance here corresponds to outputting the "no boundary" class at all locations, meaning that the false alarm rate will be zero, and the miss rate will be 1. Since the TDT metric assigns a weight of 0.7 to false alarms, and 0.3 to misses, chance in this case will be 0.3.

As shown, the error rate for the prosody model alone is lower than that for the language model. Furthermore, combining the models yields a significant improvement. Using the combined model, the error rate decreased by 27.3% relative to the language model, for the correct words, and by 24.2% for recognized words.

3.3.3 Performance without F0 features

As in the earlier case of Broadcast News sentence segmentation, since this task made use of F0 features, we asked how well it would fare without any F0 features. The experiments were conducted only for true words, since as shown previously in Table 5, results are similar to those for recognized words. Results, as shown in Table 6, indicate a significant degradation in performance when the F0 features are removed.

3.4 Comparisons of error reduction across conditions

To compare performance of the prosodic, language, and combined models directly across tasks and corpora, it is necessary to normalize over three sources of variation. First, our conditions differ in chance performance (since the percentage of boundaries that correspond to a sentence or topic change differ across tasks and copora). Second, the upper bound on accuracy in the case of imperfect word recognition depends on both the word error rate of the recognizer for the corpus, and the task. Third, the (standard) metric we have used to evaluate topic boundary detection differs from the straight accuracy metric used to assess sentence boundary detection.

A meaningful metric for comparing results directly across tasks is the percentage of the chance error that remains after application of the modeling. This measure takes into account the different chance values, as well as the ceiling effect on accuracy due to recognition errors. Thus, a model with a score of 1.0 does no better than chance for that task, since 100% of the error associated with chance performance remains after the modeling. A model with a score close to 0.0 is a nearly "perfect" model, since it eliminates nearly all the chance error. Note that in the case of recognized words, this amounts to an error rate at the lower bound rather than at zero.

In Fig. 9, performance on the relative error met-

ric is plotted by task/corpus, reliability of word cues (ASR or reference transcript), and model. In the case of the combined model, the plotted value reflects performance for whichever of the two combination approaches (HMM or interpolation) yielded best results for that condition.

Useful cross-condition comparisons can be summarized. For all tasks and as expected, performance suffers for recognized words compared with transcribed words. For the sentence segmentation tasks, the prosodic model degrades less on recognized words relative to true words than the word-based models. The topic segmentation results based on language model information show remarkable robustness to recognition errors—much more so than sentence segmentation. This can be noted by comparing the large loss in performance from reference to ASR word cues for the language model in the two sentence tasks, to the identical performance of reference and ASR words in the case of the topic task. The pattern of results can be attributed to the different character of the language model used. Sentence segmentation uses a higher-order N-gram that is sensitive to specific words around a potential boundary, whereas topic segmentation is based on bag-of-words models that are inherently robust to individual word errors.

Another important finding made visible in Fig. 9 is that the performance of the language model alone on Switchboard transcriptions is unusually good, when compared with the performance of the language model alone for all other conditions (including the corresponding condition for Broadcast News). This advantage for Switchboard completely disappears on recognized words. While researchers typically have found Switchboard a difficult corpus to process, in the case of sentence segmentation on true words it is just the opposite—atypically easy. Thus, previous work on automatic segmentation on Switchboard transcripts (Stolcke and Shriberg, 1996) is likely to overestimate success for other corpora. The Switchboard sentence segmentation advantage is due in large part to the high rate of a small number of words that occur sentence-initially (especially "I", discourse markers, backchannels, coordinating conjunctions, and disfluencies).

Finally, a potentially interesting pattern can be seen when comparing the two alternative model combination approaches (integrated HMM, or interpo-

lation) for the sentence segmentation task.[9] Only the best-performing model combination approach for each condition (ASR or reference words) is noted in Fig. 9; however, the complete set of results is inferrable from Tables 2 and 4. As indicated in the tables, the same general pattern obtained for both corpora. The integrated HMM was the better approach on true words, but it fared relatively poorly on recognized words. The posterior interpolation, on the other hand, yielded smaller, but consistent improvements over the individual knowledge sources on both true and recognized words. The pattern deserves further study, but one possible explanation is that the integrated HMM approach as we have implemented it assumes that the prosodic features are independent of the words. Recognition errors, however, will tend to affect both words (by definition) and prosodic features through incorrect alignments. This will cause the two types of observations to be correlated, violating the independence assumption.

3.5 General discussion and future work

There are a number of ways in which the studies just described could be improved and extended in future work. One issue for the prosodic modeling is that currently, all of our features come from a small window around the potential boundary. It is possible that prosodic properties spanning a longer range could convey additional useful information. A second likely source of improvement would be to utilize information about lexical stress and syllable structure in defining features (for example, to better predict the domain of prefinal lengthening). Third, additional features should be investigated; in particular it would be worthwhile to examine energy-related features if effective normalization of channel and speaker characteristics could be achieved. Fourth, our decision tree models might be improved by using alternative algorithms to induce combinations of our basic input features. This could result in smaller and/or better-performing trees. Finally, as mentioned earlier, testing on recognized words involved a fundamental mismatch with respect to model training, where only true words were used. This mismatch worked against us, since the (fair) testing on recognized words used prosodic models that

---

[9]The interpolated model combination is not possible for topic segmentation, as explained earlier.
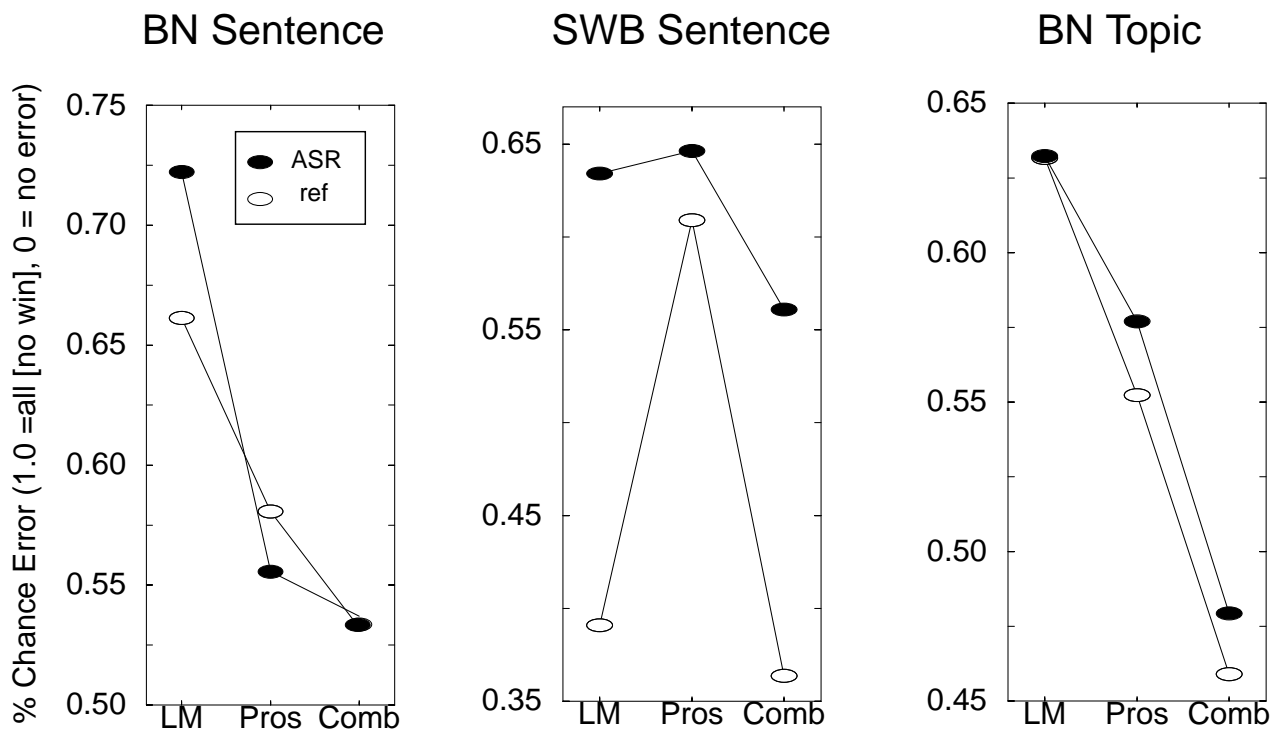
Fig. 9: Percentage of chance error remaining after application of model (allows performance to be directly compared across tasks). BN=Broadcast News, SWB=Switchboard, ASR=1-best recognition hypothesis, ref=transcribed words, LM=language model only, Pros=prosody model only, Comb=combination of language and prosody models.

had been optimized for alignments from true words. Full retraining of all model components on recognized words would be an ideal (albeit presently expensive) solution to this problem.

Comparisons between the two speech styles in terms of prosodic feature usage would benefit from a study in which factors such as speaker overlap in train and test data, and the sound quality of recordings, are more closely controlled across corpora. As noted earlier, Broadcast News had an advantage over Switchboard in terms of speaker consistency, since as is typical in speech recognition evaluations on news speech, it included speaker overlap in training and testing. This factor may have contributed to more robust performance for features dependent on good speaker normalization—particularly for the F0 features, which used an estimate of the speaker's baseline pitch. It is also not yet clear to what extent performance for certain features is affected by factors such as recording quality and bandwidth, versus aspects of the speaking style itself. For example, it is possible that a high-quality, full-bandwidth recording of Switchboard-style speech would show a greater use of prosodic features than found here.

An added area for further study is to adapt prosodic or language models to the local context. For example, Broadcast News exhibits an interesting variety of shows, speakers, speaking styles, and acoustic conditions. Our current models contain only very minimal conditioning on these local properties. However, we have found in other work that tuning the topic segmenter to the type of broadcast show provided significant improvement (Tür et al., 2000). The sentence segmentation task could also benefit from explicit modeling of speaking style. For example, our results show that both lexical and prosodic sentence segmentation cues differ substantially between spontaneous and planned speech. Finally, results might be improved by taking advantage of speaker-specific information (i.e. behaviors or tendencies beyond those accounted for by the speaker-specific normalizations included in the prosodic modeling). Initial experiments suggest we did not have enough training data per speaker available for an investigation of speaker-specific modeling; however, this could be made possible through additional data or the use of smoothing approaches to adapt global models to speaker-specific ones.

More sophisticated model combination approaches that explicitly model interactions of lexical and prosodic features offer much promise for future improvements. Two candidate approaches are the decision trees based on unsupervised hierarchical word clustering of (Heeman and Allen, 1997), and the feature selection approach for exponential models (Beeferman et al., 1999). As shown in Stolcke and Shriberg (1996) and similar to Heeman and Allen (1997), it is likely that the performance of our segmentation language models would be improved by moving to an approach based on word classes.

Finally, the approach developed here could be extended to other languages, as well as to other tasks. As noted in Section 1.3, prosody is used across languages to convey information units (e.g., (Vaissière, 1983), among others). While there is broad variation across languages in the manner in which information related to item salience (accentuation and prominence) is conveyed, there are similarities in many of the features used to convey boundaries. Such universals include pausing, pitch declination (gradual lowering of F0 valleys throughout both sentences and paragraphs), and amplitude and F0 resets at the beginnings of major units. One could thus potentially extend this approach to a new language. The prosodic features would differ, but it is expected that for many languages, similar basic raw features of pausing, duration, and pitch can be effective in segmentation tasks. In a similar vein, although prosodic features depend on the type of events one is trying to detect, the general approach could be extended to tasks beyond sentence and topic segmentation (see, for example, Hakkani-Tür et al., 1999; Shriberg et al., 1998).

4 Summary and conclusion

We have studied the use of prosodic information for sentence and topic segmentation, both of which are important tasks for information extraction and archival applications. Prosodic features reflecting pause durations, suprasegmental durations, and pitch contours were automatically extracted, regularized, and normalized. They required no hand-labeling of prosody; rather, they were based solely on time alignment information (either from a forced alignment or from recognition hypotheses).

The features were used as inputs to a decision

27

tree model, which predicted the appropriate segment boundary type at each inter-word boundary. We compared the performance of these prosodic predictors to that of statistical language models capturing lexical correlates of segment boundaries, as well as to combined models integrating both lexical and prosodic information. Two knowledge source integration approaches were investigated: one based on interpolating posterior probability estimators, and the other using a combined HMM that emitted both lexical and prosodic observations.

Results showed that on Broadcast News the prosodic model alone performed as well as (or even better than) purely word-based statistical language models, for both true and automatically recognized words. The prosodic model achieved comparable performance with significantly less training data, and often degraded less due to recognition errors. Furthermore, for all tasks and corpora, we obtained a significant improvement over word-only models using one or both of our combined models. Interestingly, the integrated HMM worked best on transcribed words, while the posterior interpolation approach was much more robust in the case of recognized words.

Analysis of the prosodic decision trees revealed that the models capture language-independent boundary indicators described in the literature, such as pre-boundary lengthening, boundary tones, and pitch resets. Consistent with descriptive work, larger breaks such as topics, showed features similar to those of sentence breaks, but with more pronounced pause and intonation patterns. Feature usage, however, was corpus dependent. While features such as pauses were heavily used in both corpora, we found that pitch is a highly informative feature in Broadcast News, whereas duration and word cues dominated in Switchboard. We conclude that prosody provides rich and complementary information to lexical information for the detection of sentence and topic boundaries in different speech styles, and that it can therefore play an important role in the automatic segmentation of spoken language.

Acknowledgements

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop* (pp. 194–218). Lansdowne, VA: Morgan Kaufmann.

Bahl, L. R., Brown, P. F., de Souza, P. V., and Mercer, R. L. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(7), 1001–1008.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions in Markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, *34*(1-3), 177–210. (Special Issue on Natural Language Learning)

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Pacific Grove, CA: Wadsworth and Brooks.

Brown, G., Currie, K., and Kenworthy, J. (1980). *Questions of Intonation.* London: Croom Helm.

Bruce, G. (1982). Textual aspects of prosody in Swedish. *Phonetica*, *39*, 274–287.

Buntine, W., and Caruana, R. (1992). *Introduction to IND Version 2.1 and Recursive Partitioning.* Moffett Field, CA.

Cieri, C., Graff, D., Liberman, M., Martey, N., and Strassell, S. (1999). The TDT-2 text and speech corpus. In *Proceedings DARPA Broadcast News Workshop* (pp. 57–60). Herndon, VA: Morgan Kaufmann.

Dermatas, E., and Kokkinakis, G. (1995). Automatic stochastic tagging of natural language texts. *Computational Linguistics*, *21*(2), 137–163.

Digalakis, V., and Murveit, H. (1994). GENONES: An algorithm for optimizing the degree of tying in a large vocabulary hidden Markov model based speech recognizer. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 537–540). Adelaide, Australia.

Doddington, G. (1998). The Topic Detection and Tracking Phase 2 (TDT2) evaluation plan. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop* (pp. 223–229). Lansdowne, VA:

Morgan Kaufmann. (Revised version available from http://www.nist.gov/speech/tdt98/tdt98.htm)

ESPS Version 5.0 Programs Manual. (1993). Washington, D.C.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCH-BOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 517–520). San Francisco.

Graff, D. (1997). The 1996 Broadcast News speech and language-model corpus. In *Proceedings DARPA Speech Recognition Workshop* (pp. 11–14). Chantilly, VA: Morgan Kaufmann.

Grosz, B., and Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, and G. E. Wiebe (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (Vol. 1, pp. 429–432). Banff, Canada.

Hakkani-Tür, D., Tür, G., Stolcke, A., and Shriberg, E. (1999). Combining words and prosody for information extraction from speech. In *Proceedings of the 6th European Conference on Speech Communication and Technology* (Vol. 5, pp. 1991–1994). Budapest.

Hearst, M. A. (1997). TexTiling: Segmenting text info multi-paragraph subtopic passages. *Computational Linguistics*, *23*(1), 33–64.

Heeman, P., and Allen, J. (1997). Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics.* Madrid.

Hirschberg, J., and Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (pp. 286–293). Santa Cruz, CA.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *35*(3), 400–401.

Koopmans-van Beinum, F. J., and van Donzel, M. E. (1996). Relationship between discourse structure and dynamic speech rate. In H. T. Bunnell and W. Idsardi (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (Vol. 3, pp. 1724–1727). Philadelphia.

Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (pp. 286–288). Ohio State University, Columbus, Ohio.

Kubala, F., Schwartz, R., Stone, R., and Weischedel, R. (1998). Named entity extraction from speech. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop* (pp. 287–292). Lansdowne, VA: Morgan Kaufmann.

Lehiste, I. (1979). Perception of sentence and paragraph boundaries. In B. Lindblom and S. Öhman (Eds.), *Frontiers of Speech Communication Research* (pp. 191–201). London: Academic.

Lehiste, I. (1980). The phonetic structure of paragraphs. In S. Nooteboom and A. Cohen (Eds.), *Structure and Process in Speech Perception* (pp. 195–206). Berlin: Springer.

Liu, D., and Kubala, F. (1999). Fast speaker change detection for Broadcast News transcription and indexing. In *Proceedings of the 6th European Conference on Speech Communication and Technology* (Vol. 3, pp. 1031–1034). Budapest.

Meteer, M., Taylor, A., MacIntyre, R., and Iyer, R. (1995). *Dysfluency Annotation Stylebook for the Switchboard Corpus.* Distributed by LDC, ftp://ftp.cis.upenn.edu-/pub/treebank/swbd/doc/DFL-book.ps. (Revised June 1995 by Ann Taylor.)

Nakajima, S., and Tsukada, H. (1997). Prosodic features of utterances in task-oriented dialogues. In Y. Sagisaka, N. Campbell, and N. Higuchi (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 81–94). New York: Springer.

National Institute for Standards and Technology. (1997). *Conversational Speech Recognition Workshop DARPA Hub-5E Evaluation.* Baltimore, MD.

National Institute for Standards and Technology. (1999). *LVCSR Hub-5 Workshop.* Linthicum Heights, MD.

Palmer, D. D., and Hearst, M. A. (1997). Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, *23*(2), 241–267.

Przybocki, M. A., and Martin, A. F. (1999). The 1999 NIST speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking. In *Proceedings of the 6th European Conference on Speech Communication and Technology* (Vol. 5, pp. 2215–2218). Budapest.

Sankar, A., Weng, F., Rivlin, Z., Stolcke, A., and Gadde, R. R. (1998). The development of SRI's 1997 Broadcast News transcription system. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop* (pp. 91–96). Lansdowne, VA: Morgan Kaufmann.

Shriberg, E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the XIVth International Congress on Phonetic Sciences* (pp. 619–622). San Francisco.

Shriberg, E., Bates, R., and Stolcke, A. (1997). A prosody-only decision-tree model for disfluency detection. In G. Kokkinakis, N. Fakotakis, and E. Dermatas (Eds.), *Proceedings of the 5th European Conference on Speech Communication and Technology* (Vol. 5, pp. 2383–2386). Rhodes, Greece.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, *41*(3-4), 439–487.

Silverman, K. (1987). *The Structure and Processing of Fundamental Frequency Contours.* Unpublished doctoral dissertation, Cambridge University, Cambridge, U.K.

Sluijter, A., and Terken, J. (1994). Beyond sentence prosody: Paragraph intonation in Dutch. *Phonetica*, *50*, 180–188.

29

Sönmez, K., Shriberg, E., Heck, L., and Weintraub, M. (1998). Modeling dynamic prosodic variation for speaker verification. In R. H. Mannell and J. Robert-Ribes (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (Vol. 7, pp. 3189–3192). Sydney: Australian Speech Science and Technology Association.

Sönmez, K., Heck, L., and Weintraub, M. (1999). Speaker tracking and detection with multiple speakers. In *Proceedings of the 6th European Conference on Speech Communication and Technology* (Vol. 5, pp. 2219–2222). Budapest.

Stolcke, A., and Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. In H. T. Bunnell and W. Idsardi (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (Vol. 2, pp. 1005–1008). Philadelphia.

Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauché, M., Tür, G., and Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In R. H. Mannell and J. Robert-Ribes (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (Vol. 5, pp. 2247–2250). Sydney: Australian Speech Science and Technology Association.

Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tür, G., Rivlin, Z., and Sönmez, K. (1999). Combining words and speech prosody for automatic topic segmentation. In *Proceedings DARPA Broadcast News Workshop* (pp. 61–64). Herndon, VA: Morgan Kaufmann.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, *101*, 514–521.

Swerts, M., and Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, *37*, 21–43.

Swerts, M., and Ostendorf, M. (1997). Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, *22*(1), 25–41.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Klein and K. K. Paliwal (Eds.), *Speech Coding and Synthesis.* New York: Elsevier.

Thorsen, N. G. (1985). Intonation and text in Standard Dutch. *Journal of the Acoustical Society of America*, *77*, 1205–1216.

Tür, G., Hakkani-Tür, D., Stolcke, A., and Shriberg, E. (2000). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, to appear.

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler and D. R. Ladd (Eds.), *Prosody: Models and Measurements* (pp. 53–66). Berlin: Springer.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*, 260–269.

Yamron, J., Carp, I., Gillick, L., Lowe, S., and van Mulbregt, P. (1998). A hidden Markov model approach to text segmentation and event tracking. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 333–336). Seattle, WA.

30