Speech Segmentation and its Impact on Spoken Document Processing

M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. G. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang and C. Wooters

Abstract

Progress in both speech and language processing has spurred efforts to support applications that rely on spoken—rather than written—language input. A key challenge in moving from text-based documents to such "spoken documents" is that spoken language lacks explicit punctuation and formatting, which can be crucial for good performance. This paper describes different levels of speech segmentation, approaches to automatically recovering segment boundary locations, and experimental results demonstrating impact on several language processing tasks. The results also show a need for optimizing segmentation for the end task rather than independently.

I. Introduction

Dramatic improvements in automatic speech recognition (ASR) technology make it now possible explore how language processing techniques designed for text can be applied to spoken language. Ever increasing collections of information are available as speech recordings, including news broadcasts, meetings, debates, lectures, hearings, oral histories, and webcasts, among other types of human-directed (vs. computer-directed) communications. Given the vast amount of audio data, and the time involved in human listening, clearly some automatic means for data processing is necessary. ASR can automatically

Manuscript received October 2007; revised December 2007. This work was supported by DARPA, contract No. HR0011-06-C-0023. Distribution is unlimited. The views herein are those of the authors and do not reflect the views of the funding agency.

M. Ostendorf is with the Department of Electrical Engineering, University of Washington. Seattle, WA, 98195 USA (email: {mo}@ee.washington.edu) Other authors are with ICSI (Favre, Hakkani-Tur, Shriberg), NYU (Grishman, Ji), University of Maryland (Harper), University of Washington (Hillard, Kahn), Columbia University (Hirschberg, Maskey, Rosenberg), University of Texas, Dallas (Liu), RWTH Aachen (Matusov, Ney), SRI International (Shriberg, Wang), and NextIT Corp. (Wooters).

transcribe (albeit imperfectly) the speech in such spoken documents into a stream of words. But to derive content of interest, one would like to be able to apply language processing techniques, such as question answering and summarization, which have traditionally been developed for written input.

In applying text-based language processing techniques to speech, it is important to acknowledge that there are many differences between the two genres, as well as within each genre, depending on the specific communication context. For example, conversational speech is generated "on the fly" and thus contains disfluencies and discourse elements used to manage turn-taking and other forms of interaction. In contrast, news broadcasts are typically read from a script, and more closely resemble written documents. Nevertheless, a shared challenge for the processing of most classes of spoken documents—as compared with text documents—is the lack of overt segmentation information. Text input typically contains punctuation and capitalization, which segments words into sentences and subsentential units. Sentences are further organized into higher-level units such as speaker quotes, paragraphs, sections, chapters, articles, and so on, via formatting. In contrast, when spoken language is processed by an automatic speech recognizer, the output is simply an unannotated stream of words, as shown in the example below. Human listeners can easily segment such spoken input, arriving at the formatted version below. To do so they can draw on sophisticated syntactic, semantic, acoustic, prosodic, pragmatic, and discourse knowledge—not all of which are fully understood.

Unformatted Word Transcripts

with more american firepower being considered for the persian gulf defense secretary cohen today issued by far the administration's toughest criticism of the u. n. security council without mentioning russia or china by name cohen took dead aim at their reluctance to get tough with iraq frankly i find it uh incredibly hard to accept the proposition that in the face of saddam's uh actions that uh members of the security council cannot bring themselves to declare that this is a fundamental or material breach uh of uh conduct on his part i think it challenges the credibility of the security council in europe today secretary of state albright trying to gather support for tougher measures was told by the british and french that before they will join the u. s. in using force they insist the security council pass yet another resolution british prime minister blair said if saddam hussein then does not comply the only option to enforce the security council's will is military action

Formatted transcripts

Reporter: With more American firepower being considered for the Persian Gulf, defense secretary Cohen today issued by far the administration's toughest criticism of the U.N. Security Council. Without mentioning Russia or China by name, Cohen took dead aim at their reluctance to get tough with Iraq.

Cohen: Frankly I find it incredibly hard to accept the proposition that in the face of Saddam's actions that members of the Security Council cannot bring themselves to declare that this is a fundamental or material breach of conduct on his part. I think it challenges the credibility of the Security Council.

Reporter: In Europe today, Secretary of State Albright trying to gather support for tougher measures was told by the British and French that before they will join the U.S. in using force they insist the security council pass yet another resolution. British Prime Minister Blair said if Saddam Hussein then does not comply:

Blair: The only option to enforce the security council's will is military action.

Automatic segmentation, on the other hand, is much more difficult. Nevertheless, significant progress has been made in the area of sentence segmentation by combining lexical information from a word recognizer, with spectral and prosodic cues. Lexical sequence information provides cues related to syntactic and semantic constraints, and is thus helpful in finding sentence and clause boundaries. For example, a sentence in English is not likely to end with a determiner. Such cues are, however, subject to degradation from word recognition errors. Lexical cues can also be fairly domain-specific, and may thus perform poorly when training and test data come from different speaking contexts. Spectral information provides cues to speaker and show (or scene) changes, as well as to non-speech events such as laughter. Prosodic features such as fundamental frequency, duration, and energy patterns provide information about multiple types of segment boundaries. For example, pitch tends to drop before the ends of sentences, and to an even lower value at the end of a topic or paragraph-like unit. Boundaries are often accompanied by pauses and by durational lengthening of phones directly preceding the boundary.

For many language processing tasks, it is essentially impossible to process speech without some sort of segmentation. Historically, spoken language processing has assumed the availability of good sentence and document segmentation. Most initial work on problems such as speech parsing and summarizing speech were based on oracle conditions using hand-marked sentence and story boundaries. However, automatically detecting these boundaries is challenging, and several studies have demonstrated that segmentation and punctuation prediction accuracy significantly impact language processing performance

in a variety of tasks, as surveyed in Section IV. Hence, over the past decade researchers have been exploring methods for improving the accuracy of models for various levels of segmentation, showing that combining both acoustic and lexical cues provides significant benefit to detection accuracy beyond a naive pause-based segmentation. More importantly, as will be shown here for a variety of language processing tasks, these segmentations also lead to much better task performance than pause-based segmentations. In addition, when looking at findings over a range of tasks, it is clear that optimizing segmentation for the task is a useful strategy, i.e., the best tradeoff of recall and precision varies depending on the task.

In the remainder of the paper, we describe different types of segmentation useful for spoken document processing (Section II), outline popular methods for feature extraction and computational modeling (Section III), survey recent results in several language processing applications that demonstrate the impact of speech segmentation (Section IV), and discuss open challenges for leveraging and augmenting automatic segmentation (Section V). For historical reasons, most research has been conducted on broadcast news; the studies represented here reflect that bias. While there is less research on segmentation of conversational speech, it has particular importance because, unlike news broadcasts, communications such as meetings or telephone conversations contain information that may not appear in any other source. Hence, a few results for conversational speech are included here to highlight issues associated with the processing of different genres.

II. Types of Segmentation in Spoken Language

Sentence segmentation is of particular importance for speech understanding applications—from parsing and information extraction at the more basic level, to machine translation, summarization and question answering at the application level. Sentence boundaries are also important for aiding human readability of the output of automatic speech recognition systems [1]. As noted earlier, most work aimed at language processing of speech input was originally developed for text, and thus assumes the presence of explicit sentence boundaries in the input. Even as the amount of spoken material online increases, making spoken document processing an interesting target in its own right, models continue to be trained on text data mainly because text is available in much larger quantities than is transcribed speech. Hence, automatic recognition of sentence boundaries in speech is important for automatic language processing, as is the general problem of reducing the mismatch between text and speech.

Note that while the definition of a sentence boundary is fairly clear in written text, spontaneous speech requires conventions for phenomena not typically present in written text. Such phenomena include incomplete utterances, backchannel responses such as “uhhuh”, boundaries involved in disfluencies, and

a variety of elliptical utterances that may not include a main verb (e.g., “Five.”, as in an answer to “When should we meet?”). Nevertheless, hand-labeling efforts show fairly good agreement once such conventions are established [2], allowing for high-level comparisons with read speech or text. In studies of conversational speech, it has also been beneficial to mark boundaries of different classes of sentences, or dialog acts—including statements, questions, and backchannels. For example, dialog act boundaries rather than sentence boundaries have been widely used for multiparty meeting speech [2], [3]. Dialog act boundaries are thus essentially equivalent to sentence boundaries in such work but provide additional information on utterance function.

Sentence-level information is but one of many useful levels of structure in language, as evidenced by the additional forms of punctuation (for example, commas) often available in text. For some language analysis tasks, such as parsing and entity extraction, sub-sentence punctuation is of additional value. Language generation-based techniques such as question answering and summarization may also benefit from sub-sentence structure annotations, as would speech playback in spoken document browsing applications. However, many of these applications may benefit more from an alternative to punctuation: prosodic phrase boundaries. Speakers naturally group words into semantically coherent phrases indicated by timing and pitch cues; these prosodic phrase boundaries often coincide with major syntactic constituent boundaries but have a much flatter structure than syntax. Prosodic phrase boundaries tend to coincide with commas and semi-colons, but they also occur in other syntactically important places and thus they provide smaller (and potentially more useful) units for processing.

Segmentation above the sentence level can be useful for choosing appropriate size units, depending on genre. For example, topic segmentation may not be useful for call center data, but it is important when processing longer spoken documents, such as news broadcasts that include multiple stories or meetings that may cover multiple agenda items. Similarly, speaker tracking and possibly role or identity recognition can provide useful structure in genres with multiple speakers. Simply knowing who is speaking (even without an associated name) can improve the readability of a speech transcript when there is more than one person talking. Speaker tracking is also useful for automatic analysis of conversation or meeting dynamics. In other applications, the speaker role can provide useful information, e.g., reporter vs. soundbite speaker as in our example or caller vs. agent in a call center. When speaker identification is needed, as for attribution in question answering, it benefits from speaker tracking and role recognition. Both speaker and topic segmentation can be useful in speech recognition, for acoustic and language model adaptation, respectively.

III. Computational Modeling Techniques

Two very different types of segmentation algorithms are used: audio diarization and structural segmentation. Audio diarization aims to segment an audio recording into acoustically homogeneous regions, given only features extracted from the audio signal. Audio diarization techniques can include a variety of tasks, such as distinguishing speech from music or advertisements from news. Probably the most important example is speaker diarization, sometimes referred to as the "Who Spoke When" task, which is what this paper will focus on. The term structural segmentation is used here to include tasks that represent linguistic structure, for which algorithms leverage both acoustic and lexical cues. The task receiving the most attention thus far has been sentence segmentation, but the methods are applicable to other tasks as well, including story segmentation, comma prediction, etc. The two classes of algorithms are treated separately below, followed by a discussion of how different types of segmentation may be combined. In both cases, the algorithms have been evaluated in a range of speech genres—including broadcast news, talk shows, and conversational meeting or telephone speech—as well as multiple languages. The basic mathematical framework is essentially the same for most scenarios, but the implementation details may change, (particularly feature extraction) and some examples are given below.

A. Speaker Diarization

Much of the foundation for speaker diarization comes from speaker recognition research; some of the earliest systems were developed to support work on speaker identification in broadcast news. A driving force behind current speaker diarization research is the competitive evaluations run by the US National Institute of Standards and Technology, in which speaker diarization must be performed with little knowledge of the characteristics of the audio or of the talkers involved. The systems are evaluated in terms of Diarization Error Rate (DER), which measures the percentage of time that a system incorrectly labels the audio recording. A typical speaker diarization system may be broken down into several "standard" components [4], with the two main components being "segmentation" and "clustering." During the segmentation step (or "speaker change detection"), boundaries between acoustic events (typically due to a change of speaker) are located to create homogeneous segments of audio. Then, during clustering, all of the segments belonging to the same speaker are grouped together.

The dominant approach to segmentation involves computing a generalized log likelihood ratio at candidate boundaries, comparing the likelihoods of the data using two distributions for the subsets of data to the left and right of the boundary vs. a single distribution for the combined set. At change points, the ratio will be high. To determine the cut-off point, typically some form of regularization or prior is

used, such as the Bayesian Information Criterion, which effectively adds a penalty for increased numbers of parameters. Some of the parameters to optimize when using penalized likelihood ratios include the size of the data windows on each side of the proposed change point, the penalty term, and the form of the distributions.

The most common approach for the initial speaker clustering is hierarchical agglomerative clustering. Hierarchical agglomerative clustering typically begins with a large number of clusters which are merged pair-wise, until arriving (ideally) at a single cluster per speaker. Since the number of speakers is not known a priori, a threshold on the relative change in cluster distance is used to determine the stopping point (i.e., number of speakers). Determining the number of speakers can be difficult in applications where some speak only briefly (e.g., in news sound bites or back channels in meetings), since they tend to be clustered in with other speakers. Although there are several parameters to tune in a clustering system, the most crucial is the distance function between clusters, which impacts effectiveness of finding small clusters.

The segmentation and clustering steps are often iterated until some stopping criteria is satisfied. In subsequent passes, different models may be used, such as hidden Markov models (HMMs) for segmentation and partitioning methods in clustering. Multi-pass methods are useful for the challenge of handling speaker overlap (in meetings and talk shows) and handling noisy conditions (meetings with distant microphones, reporters calling in from the field).

The most common features used in audio diarization are cepstral features and their derivatives, as in speech recognition except without the normalization aimed at factoring out speaker and channel differences since these are exactly the types of differences that are targeted in diarization. For tasks where there are multiple microphones, such as meeting recordings, spatial information extracted from time delays between microphones is useful.

B. Structural Segmentation

As noted earlier, many types of structural segmentation (e.g., sentence boundary, comma, intonational phrase boundary, story boundary) benefit from the use of both prosodic and lexical cues. While the effectiveness of specific cues often varies depending on the type of segmentation, the mathematical frameworks and feature extraction methods are often quite similar.

1) Computational Models: In general terms, there are two basic modeling approaches used for structural segmentation: 1) detection of boundary events and 2) whole constituent modeling. The approaches can also be combined. Both models are applied after speech recognition, and take advantage of the

alignment between words (and the phones therein) and the acoustic speech signal.

Boundary event detection can be treated as a sequence tagging problem. For each word in the sequence, one must assign a boundary label to the interval between that word and the next. As such, several of the different approaches that have been applied to other tagging tasks have also been applied to boundary detection. An HMM is one of the basic models for sequence tagging problems, and HMM-like models dominated early work in speech segmentation [5]. Given the word sequence W and the prosodic features F, the most likely event sequence E is given by:

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|W,F) \approx \underset{E}{\operatorname{argmax}}\, P(W,E)P(F|E). \tag{1}$$

The transition probabilities (in $P(W,E)$) are obtained from an n-gram language model (also referred to as a hidden-event language model) characterizing the event labels and words jointly. The observation posteriors $P(F|E)$ are generated from a prosody model (e.g., a decision tree classifier or neural network). HMMs that are discriminatively trained have also been used for sentence boundary detection (e.g., [6]).

More recently, maximum entropy (Maxent) and conditional random field (CRF) classifiers have been investigated for boundary event detection [7], [8]. Unlike HMMs, Maxent and CRF approaches provide more freedom to incorporate contextual information, both using the exponential form for the conditional probabilities. For example, in Maxent:

$$P(E_i|W,F) = \frac{1}{Z_\lambda(W,F)} \exp(\sum_k \lambda_k g_k(E_i,W,F)). \tag{2}$$

A CRF models sequence information, whereas Maxent individually classifies each data sample. The features used in these modeling approaches are typically word n-grams, part-of-speech tags, and output from prosody model or directly-modeled prosodic features. The weights (λ) for the features are estimated to maximize the conditional probabilities of the training set. In [8], HMM, Maxent and CRF approaches are compared for sentence segmentation of broadcast news and conversational speech, finding that the CRF leads to the best results but by a small margin and at a higher computational cost. Because the approaches are quite different, further gains can be obtained by combining the different systems but the biggest impact on performance comes from improving the speech recognition system. Another approach that can accommodate a rich variety of features is based on combining Boostexter with a hidden-event language model [9]. Boostexter is based on the principle of boosting that combines many weak classifiers, each having a basic form of one-level decision trees using confidence-rated prediction. It has the advantage of good performance with a relatively low cost implementation. Performance differences among approaches are small compared to differences across genres, with sentence segmentation accuracy on broadcast news

being much worse than conversational speech in part because sentences in news are longer and more complex.

Whole constituent modeling considers both the beginning and the end time of a segment in determining boundary location. For many problems, the cues are local to the boundary, such as for prosodic phrase boundaries. For others, the cues extend over the entire phrase, and the whole constituent approach is preferable. Whole constituent modeling can also be useful when a maximum or minimum length constraint is needed. The challenge of modeling the constituent is that the search space is much larger than in searching for sequential boundary events based on local cues, since all possible previous segment boundaries up to the maximum must be considered. Since this is impractical for long constituents, the search space can be reduced by restricting the set of candidate boundaries. Whole constituent modeling has been used for sentence segmentation, story segmentation, and in speaker modeling where both acoustic and lexical cues are incorporated. In sentence segmentation for translation [10], an explicit sentence length model is incorporated in a log-linear combination of language model and prosody model scores. Posterior probabilities identified via boundary event detection can be included in the combination for further improvements [11]. In story segmentation, whole constituent modeling is needed for characterizing the topical coherence of sentences in the segment and extracting position-based information about lexical cue words. Again, it is useful to combine prosodic and lexical cues [12]. Recent work in diarization [13] and speaker role modeling has also investigated combining acoustic and lexical features of the whole constituent.

2) Feature Extraction: To predict the presence or absence of a boundary event between two words, the modeling approaches described above rely on various lexical, prosodic, and structural features. Lexical features typically consist of word n-grams and part-of-speech n-grams. These features are useful for identifying short utterances in spontaneous speech such as backchannels (“uhhuh”, “yeah”), for characterizing sequences of words that are unlikely to be split by a sentence boundary (“the problem”), and for representing words that are likely to start a new sentence (such as “I”). These features have different representations in different modeling approaches, for example, an n-gram LM in the HMM framework or word tuple indicators in discriminative classifier approaches.

Prosodic features reflect information about duration, pause, intonational and energy contours. Features can be extracted from automatic alignments of word and phone transcriptions with the speech signal. Duration features (such as word, pause, and phone durations) are obtained directly from alignment time marks. Since different phones (and obviously different words) have different baseline durations, duration features are typically normalized for phonetic content. In addition they may be normalized by speaker,

since speakers differ in speaking rate. The definition of pause durations in conversational speech is more tricky than in monologues, since in the former one must specify how to treat pause time during which another speaker has the floor. Useful pitch and energy features tend to capture differences across the word boundary in question, as well as slopes and normalized level of pitch or energy just before a boundary. In the case of both pitch and energy, features must be appropriately normalized (by speaker for pitch; by channel for energy). Stylization of contours often aids feature robustness.

In a number of machine learning experiments on prosodic features for sentence segmentation, e.g., [8], it has been found that different features play different roles in different genres. The approach has been to perform feature selection to arrive at useful feature sets for a given genre. In a recent study, however, it was found that with the exception of pause length (which because of turn-taking differences can differ dramatically between monologues and conversational speech), different speaking styles actually appear to share similar underlying feature distributions and separability for boundaries versus non-boundaries. (Comparisons were made using dialog act boundaries which, as noted Section II, can be viewed as equivalent to sentence boundaries when collapsed into one class.) When comparing meetings and broadcast news in English [3], F0, duration, and energy features show remarkable similarity across styles. Figure 1 provides an example for a feature measuring durational lengthening in the word before the boundary. Such results suggest, unexpectedly, that with the exception of pausing behavior, people may be marking sentence boundaries prosodically in a similar manner in both styles—even extending duration of pre-boundary words by about the same amount over non-boundaries, relatively speaking. Thus, previously-assumed genre-specific feature differences may alternatively be explained by two factors: the modeling of priors in the machine learning techniques used, and differences in pause length distributions. One might then propose that more robust cross-genre prosodic sentence segmentation models could be built via adaptation and adjustment for these two sources of variation.

In addition to lexical and prosodic features, other structural features such as speaker change and overlap information can improve the performance of a boundary detection system. Syntactic features have also been used [14] to provide phrase level constraints for sentence boundary detection. At an even higher level, in story or topic segmentation, topic-related text features from much longer windows are useful, as in TextTiling [15].

C. Multi-level Segmentation

Since the various types of segmentation are generally interdependent and since automatically detected boundaries can be errorful, soft predictions (boundary posteriors) at the different levels can be considered

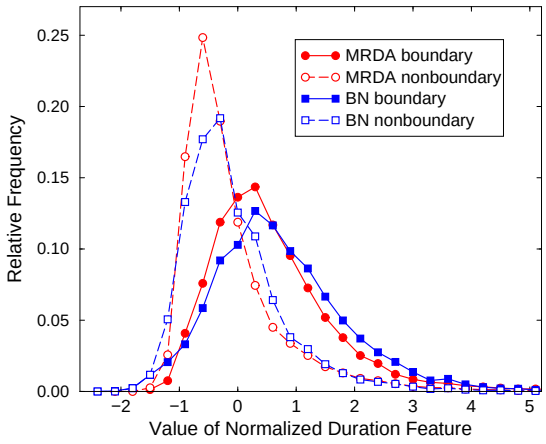

Fig. 1. Duration distributions at boundary and non-boundary events in broadcast news (BN) and meeting recordings (MRDA).

jointly to improve performance. Speaker boundaries in particular, being based purely on acoustic information, often do not align perfectly with sentence boundaries that are based on speech recognizer output. Since speaker and sentence boundaries typically coincide (except in cases of overlapping sentences, which may be seen in conversational speech), higher accuracy can be obtained by adjusting speaker boundary times to match those of nearby sentence boundaries. However, it may be more effective to include hypothesized speaker boundary scores directly into the sentence boundary detection process. At a higher level, story boundary detection also benefits from the use of soft sentence boundary decisions. In experiments on broadcast news speech [12], improved story boundary detection is achieved by considering candidate boundary points at more locations than the automatically detected sentence boundaries, either by lowering the threshold for sentence detection (e.g. from probability 0.5 to probability 0.1) or simply by considering all boundaries with a 250ms or greater length pause. Taking into consideration the higher-level information associated with story boundary detection can potentially feed back into improvements in sentence segmentation. The use of soft decisions on segment boundaries makes it possible to tune the boundary detection threshold or operating point for specific applications. Work described in the next section shows that this is indeed useful, though the best operating point varies with the different tasks.

IV. APPLICATIONS

Spoken document processing can involve a combination of several different tasks, typically starting with speech recognition and speaker segmentation, followed by some basic linguistic analysis such as part-of-speech tagging and parsing, and then involving higher level processing such as translation, information extraction and/or summarization. Automatic segmentation touches on all of these problems, but we will focus on stages after speech recognition. Often these modules are implemented in a strict pipeline, but a more tightly coupled architecture can reduce propagation of errors and improve performance. Hence, the structure of the presentation here does not imply that a pipelined architecture is required.

In the various examples here, the segmentation types used (speaker, sentence, comma, intonational phrase and story) employ the basic algorithms described in Section III. The specific variations and performance vary with genre and with the time period of the work, since this is still an area of research and the best case configurations are evolving.

Since this section describes a variety of applications, the work is evaluated with a variety of different measures, typically with standard scoring software packages or algorithms that are well documented in the literature. For brevity, detailed scoring descriptions are omitted. Most applications are scored by counting the number of correct matches to a reference. In many cases, the result is reported based on an F-score, which is the harmonic mean of precision (the percentage of detected events that are correct) and recall (the percentage of target events that are detected). For translation and summarization, scoring metrics count matches of different length word strings in the hypothesized text compared to one or more references.

A. Speaker Role and Identity Recognition

In broadcast news speech, most speech is from anchors and reporters. The remaining speech is from excerpts from quotations or interviews, sometimes referred to as “soundbites.” Detecting these soundbites and associating their speech with particular speakers is important for information extraction and attribution in question answering. This task takes as input the segmentation and speaker indexing provided by speaker diarization, and aims to determine the speaker role and name associated with each segment. Using the example in Section 1, the task is to associate the speech segments produced by Cohen and Blair with their names.

Data from the TDT4 Mandarin broadcast news has been used for soundbite segment detection and speaker name recognition [16] using a classification framework. For soundbite segment detection, each speaker turn is labeled based on the speaker’s role: anchor, reporter, or soundbite. The features used are

based on textual information, mainly word n-grams, from the current segment, the preceding and the following segments. For speaker name recognition, the approach takes advantage of the coded behavior typical of broadcast news (i.e., reporters often naming the next or previous speaker). First, hypothesized names are identified from the current and the neighboring segments, each name is classified in terms of whether or not it is the speaker's name for the target soundbite segment. The features used were words and and associated positions (e.g., in the sentence, in the segment).

The study looked at the impact of word errors and sentence segmentation errors on both the soundbite detection and name recognition tasks. For soundbite detection, the impact of segmentation errors was much greater than that of word errors, with degradation in F-measure of 20% vs. 5%, respectively. This may be due to the fact that the error rate of the recognition system is quite low (less than 10% character error rate), and the wrong sentence segmentation leads to misses of important cue words for soundbite detection. For name recognition, the opposite is true: there is no impact of segmentation error, while word errors degrade F-measure by 13%. Since many soundbite speaker names do not appear frequently in the recognizer training data, these are less reliably recognized than other words. In the pipelined implementation, errors propagate and the fully automatic system detects the soundbites and the associated names with an overall F-measure of 54%. It is likely that this result could be improved by considering multiple candidate sentence segmentations, as has been explored in other applications.

B. Tagging and Parsing

Part-of-speech (POS) tagging is the process of marking up a sequence of words with their parts of speech (e.g., noun, verb). Parsing produces a structural analysis of a word sequence with respect to a grammar. POS tagging and parsing, which are well studied and useful techniques for processing text, are now being applied to spoken language transcripts. High quality automatic sentence segmentation is fundamental for utilizing these techniques most effectively. Although a POS tagger can process word sequences that are not segmented into sentences, particularly when trained under that condition, its accuracy can be greatly improved when it is trained and evaluated on word strings segmented into sentences rather than larger segments such as stories or an entire conversation. Speech transcripts that are automatically annotated with punctuation can be tagged even more accurately. Hillard et al. [17] evaluated the impact of automatic comma prediction on POS tagging accuracy of Mandarin broadcast news speech. A Viterbi tagger trained with tag sequences from the Penn Chinese Treebank 5.2 augmented with automatically predicted commas was significantly more accurate than one trained using the same training data without punctuation.

Most natural language parsers require long word sequences to be segmented into shorter subsequences to address length-dependent complexity issues. Again parsers can be trained to process the pause-based segmentation that some speech systems use. However, since the training corpora for parsers are largely based on textual resources or employ a segmentation that is sentence-like, automatic sentence segmentation provides better matching of training and testing conditions and can improve accuracy. Until recently, research on parsing speech was done using reference transcripts hand-segmented into sentences (e.g., [18]), in part because the available parsing metrics could not measure parse performance on an input word string that did not match the yield of a reference parse. However, because parsing is a useful component technology for speech processing applications, researchers are now investigating the impact of word transcription and sentence segmentation errors on parse quality. These efforts were supported by the development of the SParseval evaluation suite [19], which can measure parse accuracy for inputs that contain word and sentence segmentation errors.

Kahn, Ostendorf, and Chelba [20] compared the effect of sentence segmentation quality on parsing of reference transcripts of conversational English. They found that parsing accuracy was greater with reference segmentation than those produced by a state-of-the-art sentence segmentation algorithm using both lexical and prosodic features, and both are significantly better than the simple pause-based segmentation of an ASR system. Using SParseval, Harper et al. [19] found similar results for ASR transcripts. In addition, they found that using ASR (vs. reference) transcripts had a slightly greater impact on parse accuracy than automatic (vs. reference) sentence segmentation. Roark et al. [14], [19] further investigated using a soft decision from sentence segmentation for parsing. Their reranking system, when optimized on two different downstream objectives—parse accuracy and sentence segmentation accuracy [19]—obtained different patterns of improvement in sentence segmentation and parse accuracy. Optimizing on sentence accuracy reduced sentence segmentation error and improved parse accuracy relative to a pause-based segmentation, but optimizing specifically for parse accuracy yield greater improvements in parsing (more than a factor of two for ASR transcripts) at some expense in sentence accuracy. When optimizing for parse accuracy, the system tended to produce shorter segments than when optimizing for sentence segmentation accuracy, i.e. trading off precision for recall. The shorter sentence-like segments also benefited a parsing language model used in speech recognition [21], leading to significant improvements in the SParseval score when word sequences and parses are chosen jointly.

C. Information Extraction

Information Extraction (IE) aims at finding semantically defined entities in documents and characterizing relations between them. Like many other text processing tasks, an IE system is often trained from text corpus with the availability of manually written punctuation. While systems for speech can be trained by removing punctuation from training data, studies have shown that there is an associated loss in performance. For example, missing commas can have a dramatic impact on IE [22], with performance loss typically bigger than that for moving from reference to ASR output (for a range of word error rates on English news). Hillard et al. [17] obtained similar results for name tagging on Mandarin broadcast news. Further, it was shown that automatic comma prediction could be used to recover half of the lost performance in experiments on text. In experiments on speech, analyses of differences in results with and without commas show cases where a comma is predicted before or after a name, which enabled the name tagger to identify a name that it had previously missed, or to correct a name boundary error.

Another study [23] confirmed these observations for English IE on speech, and found that optimizing punctuation prediction thresholds for IE performance is more effective than optimizing these thresholds separately for punctuation prediction accuracy. Favre et al. [23] focused on two types of punctuation: periods and commas, and conducted experiments using the NYU IE system [24] for a subset of TDT4 English broadcast news corpus. The results showed that removing or poorly predicting punctuation by using fixed sentence lengths adversely affects IE. Error analysis showed that punctuation errors can result in merged noun phrases or split entities. The best case performance was obtained by optimizing both comma and sentence boundary thresholds specifically for detecting entities or relations. This work suggests that punctuation should be generated differently depending on the final objective, similar to the findings for other tasks described here.

D. Machine Translation

In machine translation (MT), sentence segmentation helps provide translations with proper punctuation, but it also impacts the word choice since sentence boundaries are incorporated in the language model and the possible phrase translations. In addition, many system configurations (e.g. syntax-based statistical MT, ASR word lattice translation, rescoring and system combination algorithms for (N-best) output of one or several MT systems) require that the number of words in the input source language sentence units should not be too large (e.g. < 50 words) nor too small (e.g. > 2 words) to avoid losing context information.

The sentence length constraints motivate a constituent-based approach to sentence segmentation, in which an explicit sentence length model is included [10]. The translation application also motivates a new type of feature, introduced in [11] to characterize phrase coverage in the MT system of the words that span the candidate boundaries. The idea behind it is to make sure that word sequences with good phrasal translations will not be broken by a segment boundary. The phrase coverage feature is a bigram language model probability. Depending on whether the bigram probability is high or low, there is likely to be a good phrasal translation in the system or not, respectively. If there is a good phrasal translation, then this is probably not a good candidate for a sentence boundary.

Different sentence segmentation algorithms have been evaluated on large vocabulary Arabic-to-English and Chinese-to-English broadcast news translation tasks using the phrase-based MT system of RWTH [25]. The explicit length modeling of the whole-constituent model (using a less sophisticated prosody model and without the phrase coverage feature) did not do as well as the boundary detection approach in terms of sentence segmentation accuracy, but it did lead to better MT performance. MT performance improves by combining the two methods, but the best result was achieved by using the phrase coverage feature. The sentence boundary precision is reduced dramatically when the phrase coverage feature is used, but this does not affect the translation because the context at the erroneously inserted boundaries was not captured in MT training anyway. As in the parsing work, MT experiments have shown that a lower detection threshold is better for translation of Chinese (0.2 vs. the minimum error threshold of 0.5), favoring recall over precision. This effectively says that shorter segments are better for translation of Chinese, though average lengths depend on genre. For Arabic, longer sentences are better, and the results are less sensitive to sentence unit prediction than for Chinese-to-English translation. A separate study on Arabic-to-English translation also found that longer sentences are better, emphasizing the importance of optimizing sentence segmentation directly for translation performance [26]. Shorter sentences in Chinese are likely to help limit reordering errors, while for Arabic (which has less long distance reordering), longer segments likely provide additional context without much increased risk of reordering mistakes.

While punctuation marks predicted in ASR output can be directly translated by a MT system into target language punctuation marks, they can be also used to guide the MT process itself. In [11], automatically predicted Chinese commas were used as soft boundaries for reordering in MT search. The reordering across a comma is assumed to be highly unlikely and is penalized. This is done by modifying the lexicalized re-ordering model of the phrase-based MT system [27]. The penalty for reordering across a comma can be made dependent on the confidence with which the comma was predicted. Thus, the penalty will be smaller if the comma has a low posterior probability.

In order to test the effect of using automatically predicted commas as soft boundaries, additional experiments were performed on the Chinese-to-English task. The goal was to show that longer sentence units which capture more context can be used when reordering is constrained to sub-sentence units separated by commas. Using standard MT development scoring methods that do not require human assessment (BLEU and TER), no significant improvement was observed when using the soft boundary reordering constraints in comparison with translating shorter sentence units. However, the word order in some of the translated sentences was subjectively better when the soft boundary penalty was applied. (Punctuation is not affected, since only reordering is impacted in this implementation, so only lexical word changes are captured by the score.) It may be that intonational phrases (rather than commas) would provide better soft boundaries and/or that there are better methods for taking advantage of these cues in translation. In addition, approaches using predicted commas as features rather than constraints may be more successful.

E. Extractive Speech Summarization

Extractive speech summarization algorithms [28], [29], [30] operate by selecting segments from the source spoken documents and concatenating them to generate a summary. Generally, the speech segments extracted for summarization should be semantically meaningful and coherent stretches of speech.

Segmentation approaches currently used or proposed for extractive summarization include words, phrases, sentences, or speaker turns [28]. Choice of segmentation unit greatly influences the length and quality of the resulting summary. In experiments on English broadcast news, researchers at Columbia University explored use of intonational phrases, pause-based chunking and sentence units as alternatives for segmentation in summarization. A segment was labeled for inclusion in the summary if more than 50% of the segment was present in the human summary. Inclusion vs. exclusion was predicted automatically using a Bayesian network classifier that used only acoustic and structural features for summarization. Using the standard ROUGE summarization score, the best results were obtained with intonational phrases.

Other experiments by researchers at UT Dallas have looked at whether tuning the sentence segmentation threshold for extractive summarization could lead to improved performance. In this case, experiments were on the ICSI meeting corpus, and the inclusion vs. exclusion classifier used maximal marginal relevance with textual features only. An HMM was used for sentence segmentation, and the decision threshold was varied to provide different units for the subsequent summarization module. Unlike many of the other applications explored here, results showed that performance was stable over a large range of sentence segmentation thresholds, though this may be a consequence of the specific text features used.

V. Open Questions

In summary, the fact that most language technology used in spoken document processing is designed in large part from written text argues that speech must be made to look more like text for achieving good performance. One important challenge in this respect is speech segmentation, including sentence segmentation at a minimum, but ideally also speaker and topic segmentation for formatting and adaptation, as well as sub-sentence punctuation and/or intonational phrase prediction for higher accuracy in many applications. There are a few basic computational models that have been developed for this purpose, many of which combine lexical and acoustic cues in detecting boundaries. While these algorithms are far from perfect, in most applications they provide a much better solution than simple pause-based segmentation. Of course, there is also evidence that further improvements to segmentation algorithms would be worthwhile, though improvements to word recognition alone will provide some gains in structural segmentation performance. In addition, there are several remaining challenges for annotating speech to improve language processing.

In the various applications surveyed here, there is a consistent finding that tuning the segmentation thresholds for the application leads to significant performance improvements over using the threshold that minimizes segmentation error alone. In many cases, higher recall is more effective (i.e. shorter sentences). However, the optimal threshold varies, and in some cases longer sentences are more effective. This raises the question as to how best to meet the needs of multiple language processing modules, particularly when they must all operate on the same hypothesized word sequence. One solution is to use of a low threshold (more hypothesized boundaries) with confidences associated with the boundaries, so that different downstream modules can use their own threshold. Alternatively, it may be that the need for different thresholds reflects a need for different types of structures, including sub-sentence units such as intonational phrases or syntactic chunking. Indeed, the information extraction experiments show that the optimal sentence detection threshold interacts with the comma threshold.

Another important difference between speech and text, which was not addressed here, is the presence of disfluencies in speech. Consider the example: *I went I left the store* is a sentence containing a speech repair, where the speaker intends *I went* to be replaced by *I left*. Appropriate processing of such disfluencies poses a serious challenge, in part because they are not well modeled in textual training materials. Research has shown that automatic identification of speech repairs can significantly improve parsing [18], and that some of the structural modeling approaches used sentence segmentation can be applied to disfluency detection [8]. However, there has been relatively little research on automatic detection of disfluencies and

its use in language processing, in part because there is so little disfluency-annotated speech corpora. As language processing research increasingly turns to conversational speech tasks, such as talk shows and meetings, it will become important to address this problem.

Of course, it is important to remember that there is useful information in speech beyond what is in text. In particular, there are cues to speaker intent and information salience that are there to be mined in future applications. However, leveraging this information in language processing will require annotating large speech corpora, as well as leveraging of semi-supervised and unsupervised learning methods to avoid costly hand-labeling efforts.

REFERENCES

[1] D. Jones, E. Gibson, W. Shen, N. Granoien, M. Herzog, D. Reynolds, and C. Weinstein, "Measuring human readability of machine-generated text: Three case studies in speech recognition and machine translation," in *Proc. Inter. Conf. Acoustics, Speech, and Signal Processing*, 2005, pp. 1009–1012.

[2] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. of SIGDAL*, 2004.

[3] E. Shriberg, B. Favre, J. Fung, D. Hakkani-Tur, and S. Cuendet, "Prosodic similarities of dialog act boundaries across speaking styles," in *Linguistic Patterns in Spontaneous Speech (Language and Linguistics Monograph Series)*, S.-C. Tseng, Ed. Institute of Linguistics, Academia Sinica, Taipei, 2008.

[4] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[5] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. ICSLP*, 1996, pp. 1005–1008.

[6] M. Tomalin and P. Woodland, "Discriminatively trained Gaussian mixture models for sentence boundary detection," in *ICASSP*, 2006.

[7] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. ICSLP*, 2002, pp. 917–920.

[8] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[9] M. Zimmermann, D. Tur, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, "The ICSI+ multi-lingual sentence segmentation system," in *INTERSPEECH*, 2006.

[10] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *Proc. Inter. Workshop on Spoken Language Translation*, 2006, pp. 158–165.

[11] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tur, M. Ostendorf, and H. Ney, "Improving speech translation by automatic boundary prediction," in *Proc. Interspeech*, 2007, pp. 2449–2452.

[12] A. Rosenberg, M. Sharifi, and J. Hirschberg, "Varying input segmentation for story boundary detection in English, Arabic and Mandarin broadcast news," in *Proc. Interspeech*, 2007, pp. 2589–2592.

[13] C. Ma, P. Nguyen, and M. Mahajan, "Finding speaker identities with a conditional maximum entropy model," in *ICASSP*, vol. 4, 2007, pp. 261–264.

[14] B. Roark, Y. Liu, M. P. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L.Yung, "Reranking for sentence boundary detection in conversational speech," in *Proc. Inter. Conf. Acoustics, Speech, and Signal Processing*, 2006, pp. 545–548.

[15] M. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[16] F. Liu and Y. Liu, "Soundbite identification using reference and automatic transcripts of broadcast news speech," in *Proc. of ASRU*, 2007.

[17] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-Tür, M. Harper, M. Ostendorf, and W. Wang, "Impact of automatic comma prediction on POS/name tagging of speech," in *Proc. IEEE/ACL Workshop Spoken Language Technology*, 2006, pp. 58–61.

[18] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech," in *Proc. NAACL Conference*, 2001, pp. 118–126.

[19] M. P. Harper, B. Dorr, J. Hale, B. Roark, I. Shafran, M. Lease, Y. Liu, M. Snover, L.Yung, A. Krasnyanskaya, and R. Stewart, "2005 Johns Hopkins summer workshop final report on parsing and spoken structural event detection," Johns Hopkins University, Tech. Rep., 2005.

[20] J. G. Kahn, M. Ostendorf, and C. Chelba, "Parsing conversational speech using enhanced segmentation," in *Proc. HLT/NAACL Conference*, 2004, pp. 125–128.

[21] J. G. Kahn, D. Hillard, M. Ostendorf, and W. McNeill, "Joint optimization of parsing and word recognition with automatic segmentation," University of Washington, EE Dept., Tech. Rep., 2007.

[22] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. Ramshaw, D. Stallard, R. Schwartz, and B. Xiang, "The effects of speech recognition and punctuation on information extraction performance," in *Proc. Eurospeech*, 2005, pp. 57–60.

[23] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tur, and M. Ostendorf, "Punctuating speech for information extraction," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2008.

[24] R. Grishman, D. Westbrook, and A. Meyers, "NYU's English ACE2005 system description," in *Proc. ACE2005 Workshop*, 2005.

[25] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney, "The RWTH statistical machine translation system for the IWSLT 2006 evaluation," in *Proc. Inter. Workshop Spoken Language Translation*, 2006, pp. 103–110.

[26] S. Matsoukas, I. Bulyko, B. Xiang, K. Nguyen, R. Schwartz, and J. Makhoul, "Integrating speech recognition and machine translation," in *Proc. Inter. Conf. Acoustics, Speech, and Signal Processing*, 2007.

[27] R. Zens and H. Ney, "Discriminative reordering models for statistical machine translation," in *Proc. HLT/NAACL Workshop on Statistical Machine Translation*, 2006, pp. 55–63.

[28] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic speech summarization applied to English broadcast news speech," in *Proc. Inter. Conf. Acoustics, Speech, and Signal Processing*, 2002, pp. 9–12.

[29] X. Zhu and G. Penn, "Roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization," in *Proc of HLT/NAACL*, 2006.

[30] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "From text summarisation to style-specific summarisation for broadcast news," in *Proc. ECIR*, 2004.