Speech Segmentation and its Impact on Spoken Language Technology

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

In recent years, there has been dramatic progress in both speech and language processing, which spurs efforts to combine speech and language technologies in spoken document processing applications. However, speech is different from text in many respects, most notably in the lack of explicit punctuation and formatting. Thus, spoken document processing requires automatic segmentation to achieve good performance. This paper outlines some of the types of segmentation being used and the issues involved in computational modeling, as well as the impact on different types of language processing tasks.

I. Introduction

As large vocabulary automatic speech recognition (ASR) technology has dramatically improved in the past few years, it is now possible to explore language processing on speech sources as well as text. In other words, news broadcasts, oral histories, or recordings of a lecture or meeting can be treated as "spoken documents," which one might want to translate or use in question answering or summarization, just as written documents are now being processed. It is especially important to be able to automatically process spoken documents, because it takes much more time for a human to listen to a recording than to read a transcript. In addition, many spoken documents complement information available in written form in terms of providing more insight into opinions and emotions of the source.

Manuscript received October 2007; revised XXXX. 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), 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).

In order to apply language processing techniques to speech that have been traditionally applied to text, it is important to address the inherent differences between these two types of inputs. Although speech is similar in many ways to text (e.g., it is comprised of words that have the same meaning as in text), it also has many differences, some stemming from the fact that people use different cognitive processes when producing speech, and others stemming from the different ways in which these two methods of communication are conventionally expressed. Of course, there is a potential for automatic transcription errors in speech. However, for purposes of spoken document processing, one of the key differences between speech and text sources is that typical ASR systems have not provided punctuation or text formatting cues. Textual input typically involves words that are broken into sentences and clauses using punctuation. Sentences are further organized into chunks such as paragraphs, sections, chapters, articles, and so on. Though not traditionally exploited in word recognition systems, spoken language does include related structural cues for the listener, including pause, timing and intonational clues to sentence and topic structure, as well as indicators of speaker intention and turn-taking. These cues make it possible to turn the unformatted string of words in the example below into the formatted version that follows.

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.

Human listeners take advantage of a range of different cues in segmenting speech, including both acoustic and lexical cues. Acoustic cues include spectral cues to speaker voices and the cues associated with prosody (how a sequence of words is spoken). Acoustic-prosodic features such as fundamental frequency, duration, and energy provide information about multiple types of segment boundaries. For example, fundamental frequency tends to increase at a topic change boundary, or decrease during an aside. Segmental durations are longer and speakers often pause at major constituent boundaries. Prosody provides a valuable side channel for communicating information in speech, making spoken language richer than a simple word transcript and including cues to more than segmentation. While not addressed here, emphasis and intent are also communicated with prosodic cues. Of course, words provide powerful cues, both negative and positive. For example, a sentence is not likely to end with a determiner, and the word "now" at the start of a sentence often suggests the beginning of a new topic.

Acoustic cues are complementary to the textual cues but are often seen to play a secondary role in automatic language processing. To illustrate this point, in the example above, one could figure out many of the sentence boundaries from the text alone, though not so easily the speaker boundaries. However, the challenge would be much greater with speech recognition errors, which tends to increase the relative importance of prosodic cues. In addition, as speech becomes more informal and choices of wording and syntactic structure change, language models trained from written text are less well matched to the speech data. In general, when language cues to structure are weakened by errors and/or domain mismatch, the

same acoustic cues to structure that benefit humans can play an even more significant role in automatic processing as well.

Historically, spoken language processing has assumed the availability of good sentence and document segmentation, and most initial work on problems such as parsing and summarizing speech were based on oracle conditions using hand-marked sentence and story boundaries. In many cases, experiments were conducted in human-computer communication tasks that naturally provide isolated sentences. However, in moving to documents such as broadcast news or lectures, it is not reasonable to assume the availability of hand-marked sentence boundaries. For many language processing tasks, it is essentially impossible to process speech without some sort of segmentation, in which case a simple pause-based segmentation provides a baseline. However, several studies have demonstrated that segmentation accuracy significantly impacts language processing performance, including in parsing [1], summarization [2], and machine translation [3]. Studies on parsing [4] and information extraction [5] have shown performance degradation associated with missing sentence-internal punctuation.

Recognizing the need for automatic segmentation, many researchers have been working on this problem in the past decade for different types of segmentation. Across a range of tasks, we find that automatic segmentation driven by both acoustic and lexical cues provides significant benefit to language processing beyond a naive pause-based segmentation. In addition, we find that optimizing segmentation thresholds for language processing performance leads to better overall system performance and that the best tradeoff of recall and precision varies depending on the task.

In this paper, we will outline the different types of segmentation that seem to be useful for spoken document processing, outline popular methods for feature extraction and computational modeling, and report on recent results in several language processing applications that demonstrate the impact of speech segmentation.

II. Segmentation and Structure 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. 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 based on text data simply because text is available in huge quantities compared to hand 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. Sentence boundaries are also important for aiding human readability of the output of automatic speech recognition systems [6], as well as the accuracy and readability of the output of subsequent language processing modules such as translation and summarization.

Sentence segmentation alone provides an impoverished representation of spoken language, as evidenced by the much richer representation of structure found in written text. For some text processing tasks, such as parsing and entity extraction, sub-sentence punctuation is at least as important as sentence punctuation. 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 phrases tend to provide smaller units for processing and avoid the problem of inconsistent use of commas and other sentence-internal punctuation.

The types of segmentation above the sentence level that are useful for speech documents depends on their genre, e.g., recordings of meetings, oral histories, call center data, broadcast news. For most genres, speaker tracking and possibly role or identity recognition can provide useful structure. 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 as well as in search, it benefits from speaker tracking and role recognition. At a higher level, topic segmentation is important when processing longer spoken documents, such as meetings or news broadcasts that may cover multiple stories, in order to choose appropriately sized units for subsequent processing. 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 have been explored: segmentation purely based on acoustic coherence, as in speaker diarization, and methods that combine acoustic and word cues, as in sentence and story segmentation. These are treated separately below, followed by a discussion of how different

types of segmentation may be combined. Segmentation algorithms have also 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 different genres/languages impact details of the implementation, but the basic mathematical framework is essentially the same for most scenarios.

A. Acoustic Segmentation for Speaker Diarization

The goal of speaker diarization is to segment an audio recording into speaker-homogeneous regions, sometimes referred to as the "Who Spoke When" task. 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),[1] which measures the percentage of time that a system incorrectly labels the audio recording. Since systems are not required to know the actual names of the speakers in a recording, the evaluation software creates an optimal alignment between the system-generated labels and the true speaker labels before DER is calculated.

A typical speaker diarization system may be broken down into several "standard" components as described in [7]. The two main components of any system are "segmentation" and "clustering." During the segmentation step (sometimes referred to as "speaker change detection"), boundaries between acoustic events (typically 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 speaker segmentation involves the use of the Bayesian Information Criterion (BIC) [8] in some form or other. BIC operates by comparing two models using two windows of data on either side of a proposed change point. In the first model, all of the data from the both windows is modeled by a single distribution. In the second model, the two sets of data are modeled by two distinct distributions. The final results is given by a penalized likelihood-ratio test between the two models. There are many parameters to optimize when using BIC, including the size of the data windows on each side of the proposed change point, the penalty term, and the form of the distributions.

The second main component of a speaker diarization system is the clustering component. The most common approach for the initial speaker clustering is hierarchical agglomerative clustering. Hierarchical

[1] `http://www.nist.gov/speech/tests/rt/rt2007/`.

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 (as in a sound bite in news or simple agreements in meetings), since they tend to be clustered in with other speakers. Again, there are many parameters to be tuned in a clustering system, but 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 also 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).

B. Combining Acoustic and Lexical Features

Many types of segmentation problems (e.g., sentence boundary, comma, intonational phrase boundary, story boundary) benefit from the use of both acoustic features (such as timing, pauses and fundamental frequency) as well as wording cues. While the specific cues of interest vary across the different problems, the general mathematical frameworks and types of features used for detection apply to most examples.

1) Computational Models: In general terms, there are two basic approaches to segmentation – detection of boundary events and whole constituent modeling – which can also be combined. In both cases, the models are applied after speech recognition and take advantage of time alignments of words (and the phones therein) to the speech signal.

Boundary event detection is essentially a sequence tagging problem: for each word in the sequence, assign a boundary label to the interval between that word and the next. Different modeling approaches have been investigated for these tagging tasks, including variants of the hidden Markov model (HMM) and discriminant models based on maximum entropy (Maxent) and conditional random fields (CRFs). We briefly review these below; details on their application to sentence segmentation can be found in [9].

An HMM is one of the basic models for sequence tagging problems, and HMM-like models dominated early work in speech segmentation [10]. Given the word sequence W and the prosodic features F, the most likely event sequence E is given by:

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

The transition probabilities (in $P(W, E)$) are obtained from an N-gram language model characterizing the event labels and words jointly, which is sometimes referred to as a hidden-event language model. The observation posteriors $P(F|E)$ are generated from a prosody model, e.g., a decision tree classifier or neural network. Recently, other discriminative classifiers have been integrated into this framework, such as Boostexter, and additional lexical cues are included in the prediction. Boostexter [11] 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.

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:

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

The difference between the two approaches is in that CRF models sequence information, whereas Maxent does individual classification for each data sample. The weights (λ) for the features are estimated to maximize the conditional probabilities of the training set ($P(C|D)$), using estimation methods such as limited memory BFGS. The features used in these modeling approaches typically are N-gram of words, part-of-speech tags, and output from the prosody model or directly prosodic features.

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 intonational phrase boundaries. For others, the cues extend over the entire phrase, and the whole constituent approach is preferable. For example, in story segmentation (like speaker diarization), there is often an assumption that all sentences within a story are topically coherent. Whole constituent modeling can also be useful in cases where a maximum or minimum length constraint is needed, as in sentence segmentation for translation [3] where an explicit sentence length model is incorporated in a log-linear combination of language model and prosody model scores. In a similar way, constituent methods can easily incorporate the posterior probabilities identified via boundary event detection, as in [12]. The challenge of this approach is that the search space is much larger, since all possible previous segment boundaries up to the maximum sentence length (e.g. 40 or 50 words) must be considered.

2) Feature Extraction: The modeling approaches described above rely on various lexical, prosodic and structural features to predict the presence or absence of a boundary event between two words. Lexical features typically consist of word n-grams and part-of-speech n-grams. These features are very useful for identifying short utterances in spontaneous speech like backchannels (“okay”, “uhhuh”, “yeah”), for characterizing sequences of words that are unlikely to be split by a sentence boundary (“of the”), and

representing words that are likely to start a new sentence (such as “I”). These features have different representations in different modeling approaches, for example, the N-gram LM in the HMM framework or word tuple indicators in discriminative approaches. Syntactic features can also be used [13]. Since parsing is significantly impacted by sentence boundary detection accuracy, syntactic features may be most useful for sub-sentence constituents. At a higher level, in story or topic segmentation, topic-related text features are useful, as in TextTiling [14].

Prosodic features reflect information about temporal effects, as well as intonational and energy contours. They provide additional information complementary to the textual cues for event detection. For example, they can model phenomena that occur at sentence boundaries, such as long pauses, lengthening of word-final phonemes, and fundamental frequency (F0) changes. The prosodic features for inter-word boundary event detection are automatically extracted from the words, phonetic alignments of the transcription, and the speech signal. Duration features (such as word, pause, and phone durations) are normalized by overall phone duration statistics and speaker-specific statistics. To obtain intonation features, F0 tracks are extracted from the speech signal and then post-processed to obtain stylized contours [15], from which F0 features are extracted. Examples of intonation features are the distance from the average F0 in the word to the speaker’s F0 floor and the change in the average stylized contour across a word boundary. Similar processing is performed to obtain energy features. Note that features are also derived by comparing the two sides of a boundary, in order to model discourse continuity and better detect the occurrence of a structural event. In addition to lexical and prosodic features, we also incorporate other structural features such as speaker change and overlap information.

In studies of feature selection for sentence segmentation, different prosodic features are selected depending on the corpus and speaking style. However, different speaking styles actually share similar underlying feature distributions and separability for boundaries versus non-boundaries. In a study comparing meetings and broadcast news in English [16], F0, duration, and energy features show remarkable similarity across styles; see Figure 1 for duration lengthening in the rhyme (vowel plus following consonants) of the last syllable of the word before the boundary. This suggests that people are marking sentence boundaries prosodically in the same manner in both styles (excluding pause, which behaves differently). They are even extending duration of pre-boundary words by about the same amount over non-boundaries, relatively speaking. This suggests that more robust cross-genre prosodic sentence segmentation models could be built via adaptation and adjustment for difference in class priors.

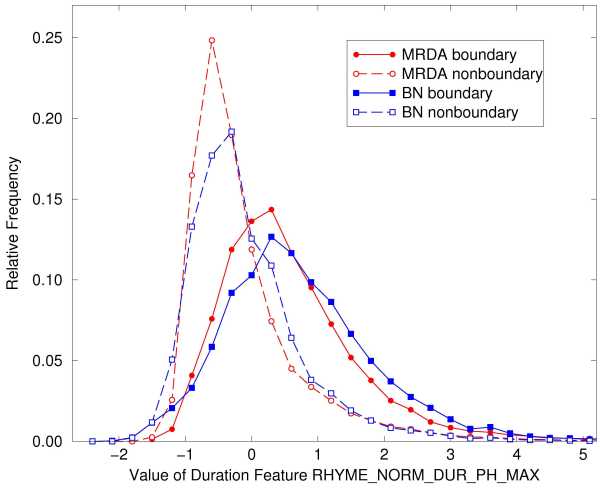

Fig. 1. Duration distributions at boundary and non-boundary events in broadcast news (BN) and meeting recordings (MRDA).

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 also be examined 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 the speaker boundary times to match 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 [17], 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 feedback 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. As for story boundary detection, we will see in the

next section that tuning the detection threshold leads to performance gains in a variety of applications, though the best operating point varies with the different tasks.

IV. Applications

A. Tagging and Parsing

Part-of-speech (POS) tagging and parsing, which are well studied and useful techniques for processing text, are now being applied to speech transcripts. POS tagging is the process of marking up a sequence of words with their parts of speech (e.g., noun, verb), and parsing produces a structural analysis of a word sequence with respect to a grammar. Tagging and parsing can be used in a variety speech applications; however, high quality automatic sentence segmentation is fundamental for utilizing these techniques most effectively. Tagging and parsing features can, in turn, be used to improve sentence segmentation accuracy.

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 conversation sides. Speech transcripts that are automatically annotated with punctuation can be tagged even more accurately. Hillard et al. [18] 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 and caesuras was significantly more accurate than one trained using the same training data without punctuation.

Most natural language parsers require words to be segmented into short segments due to their $O(n^3)$ running time. Speech systems produce transcripts for segments of speech (usually pause-based), but these often do not correspond to a sentence in text. Although parsers could be trained to process segments other than textual sentences, the training corpora for parsers are largely based on textual resources or employ a segmentation that is sentence-like (e.g., the SU) that can be reliably annotated. Hence, accurate automatic sentence segmentation is important for ensuring that the training and testing conditions of a statistical parser are well matched.

The earliest research on parsing speech (e.g., [19]) was done using gold standard transcripts segmented into sentences, 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 gold standard parse. However, because parsing is a useful component technology for speech processing applications, researchers are now investigating the impact of word and sentence segmentation errors on parse quality. These efforts were supported by the development of the SParseval evaluation suite [20], [21], which can measure parse accuracy for inputs that

contain word and sentence segmentation errors. Kahn, Ostendorf, and Chelba [1] compared the effect of sentence segmentation quality on parsing accuracy of gold standard transcripts of conversational English. They found that parsing accuracy was greater with gold standard segmentations than those produced by a state-of-the-art sentence segmentation algorithm using both lexical and prosodic features, and that the state-of-the-art-produced segments supported more accurate parsing than than the simple pause-based segmentations of an ASR system. Using SParseval, Harper et al. [20] found that parsers that process gold standard transcripts with gold standard sentence boundaries are more accurate than those produced based on ASR transcripts with automatic sentence segmentation. Although both are critical to accurate speech processing, transcript accuracy had a slightly greater impact on parse accuracy than sentence segmentation accuracy.

POS tagging- and parsing-derived features can be used to enhance the accuracy of automatic sentence segmentation. POS tags and features derived from them improve the accuracy of the ICSI sentence segmentation system (1-2% improvement in F-measure for Mandarin and English broadcast news and Mandarin broadcast conversation). Using a reranking approach on English conversational sentence hypotheses, Roark et al. [13] obtained reductions in sentence segmentation error by utilizing syntactic features. Their reranking system, when optimized on two different downstream objectives [20]– parse accuracy and sentence segmentation accuracy– obtained different patterns of improvement in sentence segmentation and parse accuracy. Optimizing on sentence accuracy enhanced both sentence and parse accuracy; however, optimizing specifically for parse accuracy yielded additional improvements in parsing accuracy, but at the expense of sentence segmentation accuracy. Interestingly, when optimizing for parse accuracy, the system tended to produce shorter word segments than when optimizing for sentence segmentation accuracy.

B. Speech Recognition

Since sentence segmentation has a significant impact on parsing, one would expect parsing language models for speech recognition to be sensitive to segmentation as well, particularly if the end goal is to have a parsed representation of the recognized speech. In speech recognition experiments with a parser as a discriminative language model vs. joint selection of the parse and word sequence, Kahn et al. [22] find that segmentation does impact the effectiveness of the parsing language model for minimizing word error rate, but the impact is much greater when the parse of the recognized word sequence is taken into account (8% reduction in WER vs. 20% reduction in the SParseval error, 1-F). In addition, the sentence boundary detection threshold that optimizes parser performance (from the experiments described above)

also gives better word recognition performance than that which optimizes sentence detection accuracy. Further improvements are obtained using the oracle segmentation, indicating the additional research in sentence boundary detection would be useful. In these experiments, the types of errors that a parsing language model corrected include, among other things, lost pronouns. These short words are easily confusable and have a high error but are important to recover for entity recognition and attribution.

Other types of segmentation that might impact recognition are speaker diarization and topic segmentation, for acoustic and language modeling adaptation, respectively. The basic clustering technology behind speaker diarization has long been used in speaker adaptation, particularly in recognition of broadcast news, but advances in diarization have not led to gains in ASR, probably because for ASR purposes it is sometimes advantageous to split the speech from a single speaker when they are recorded in different conditions (field vs. studio reporting, for example) and group speakers with very little speech. Topic segmentation has not yet been extensively explored for language model adaptation, since the most effective adaptation techniques have been mixture models that work well with sentence-level topic decisions. However, it offers the potential for gains in future work with new models.

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, IE presupposes the availability of punctuation. Researchers from BBN showed that missing commas can have a dramatic impact on IE [5], with performance losses typically bigger than that for moving from reference to automatic sentence segmentation (for a range of word error rates on English news). In our study we confirmed these results for both Mandarin and English IE on speech, and further looked at whether IE performance can be improved by generating punctuation with a view toward IE performance instead of the accuracy of the punctuation prediction itself.

In [18], we evaluated the effect of automatic comma prediction on Mandarin name tagging for 881 Mandarin broadcast news sentences. Compared to ASR output without commas, the comma predictions changed the name tagging of 59 tokens; 44 incorrect tags were corrected, 9 correct tags were changed to incorrect ones, and 6 incorrect tags were changed to other incorrect tags. In examining the changes, we observed a number of cases where the comma predictor was able to predict a comma before a name, and this enabled the name tagger to identify a name that it had previously missed, or to correct a name boundary error.

In another recent study we focused on two types of punctuation: periods and commas, and conducted

System	Period Threshold	Comma Threshold	Entity Value	Relation Value
Ref. Punc.	-	-	46.9	20.0
Fixed Sentence Length	-	-	45.7	17.4
Opt. Punc.	0.27	0.68	46.1	17.6
Opt. Entity	0.09	0.50	**48.2**	16.9
Opt. Relation	0.21	0.28	46.1	**18.4**

TABLE I

COMPARING ENGLISH IE PERFORMANCE ON SPEECH RECOGNIZER OUTPUT WHEN PUNCTUATION IS EXTRACTED FROM REFERENCE TRANSCRIPTIONS (REF. PUNC.), INSERTED TO PROVIDE FIXED 15 WORD SENTENCE LENGTH (FIXED SENTENCE LENGTH), OPTIMIZED FOR PUNCTUATION PREDICTION F-MEASURE (OPT. PUNC.) OR OPTIMIZED IN ORDER TO IMPROVE ENTITIES (OPT. ENTITY) AND RELATIONS (OPT. RELATION).

experiments using the NYU IE system [23] for the portion of TDT4 English broadcast news corpus that overlaps with the ACE (Automatic Content Extraction) 2004 reference data. The speech was transcribed by SRI's English Broadcast News speech recognizer [24] with an estimated word error rate of 18%.

Table I presents ACE entity and relation scores[2] using different ways of inserting punctuation. The results show 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. Setting the punctuation decision thresholds to maximize punctuation performance gives better results but is sub-optimal for IE. The best case peformance is obtained by optimizing both comma and sentence boundary thresholds specifically for annotating entities or relations. This suggests that punctuation should be generated differently depending on the final objective.

D. Speaker Role and Identity Recognition

In Broadcast news speech, most of the speech is from anchors and reporters. Others are excerpts from quotations or interviewees, called "soundbites" [25]. Detecting these soundbites and their speaker names is useful for information extraction and attribution in question answering.

For soundbite segment detection, we classify the segment of each speaker turn based on the speaker's role: anchor, reporter, or soundbite. The features we used are based on textual information [26], mainly word N-grams, from the current segment, the preceding and the following segments. We also found that

[2]ACE 2004 scoring metric can be found at http://www.itl.nist.gov/iaui/894.01/tests/ace/ace04/doc/ace04-evalplanv7.pdf

using words from the first and the last sentence in the segment outperforms using all the words, and that performing a three-way classification (anchor, reporter, soundbite) and then grouping anchor and reporter segments to non-soundbites is better than a direct binary classification setup. For soundbite speaker name recognition, we first identify the name hypotheses from the current and the neighboring segments, then determine for each name whether it is the speaker's name for the soundbite segment. The features we used are words, structural information (position in the sentence, in the segment). For both of these tasks, we show results using a SVM classifier, focusing on the impact of sentence segmentation.

We use the Mandarin TDT4 data for this experiment, and human annotated speaker turn segments. There are 24 broadcast news shows in the test set, and about 114 soundbite segments. The results of soundbite detection, soundbite speaker name recognition, and the pipeline system combining both are presented in Table II. "REF" means the human transcripts and human annotated sentences. "ASR_ASB" means using ASR output and automatic sentence segmentation results. "ASR_RSB" is obtained by aligning reference sentence boundaries in the human transcripts to the ASR words.

Test set	soundbite detection			soundbite name recognition	soundbite detection and name recognition		
	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
ASR_ASB	68	46.79	55.44	59.5	51.8	34.5	41.4
ASR_RSB	73.96	65.14	69.27	59.5	55.9	45.2	50
REF	74.31	71.05	72.65	68.6	51.6	57	54.1

TABLE II

SOUNDBITE DETECTION AND NAME RECOGNITION RESULTS.

We observe that speech recognition errors hurt the system performance (comparing REF and ASR conditions). Using automatic sentence boundary detection degrades performance even more for soundbite detection (comparing RSB and ASB). In particular, there is a significant decrease of the recall rate when using automatic sentence boundary hypotheses. It might be because that the error rate of the ASR output we used is quite low for this BN data, and the wrong sentence segmentation leads to misses of important cue words for soundbite detection. Different from the soundbite detection module, we can see that using automatically detected sentence boundaries has a negligible effect on soundbite speaker name recognition accuracy. This is not very surprising, since the features used for name recognition rely less on the sentence

boundary than for soundbite detection. The pipeline system performance degrades significantly because of error propagation of all types: soundbite detection, word recognition, and sentence boundary detection.

E. Machine Translation

In machine translation (MT), it is important to produce translations of sentences or sentence-like units with proper punctuation to make them human-readable. Also, sophisticated speech translation algorithms (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) often require that the number of words in the input source language SUs 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 [3]. The translation application also motivates a new type of feature, introduced in [12] 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. The translation system used was the state-of-the-art phrase-based MT system of RWTH [27]. 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. Performance improves by combining the two methods, but the best result was achieved by using the phrase coverage feature. The 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 also shown that shorter segments are better for translation of Chinese, i.e. recall is more important than precision. However, for Arabic, longer sentences are better, and the results are less sensitive to SU prediction than for Chinese-to-English translation.

Whereas the punctuation marks predicted in the 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 [12],

automatically predicted Chinese commas are 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 [28]. The penalty for reordering across a comma can be made dependent on the confidence with which this 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, we performed additional experiments on the Chinese-to-English task. The goal was to show that longer SUs which capture more context can be used when reordering is constrained to sub-sentence units separated by commas. Unfortunately, so far we observed no significant improvement in BLEU scores [29] when using the soft boundary reordering constraints in comparison with translating shorter SUs. Nevertheless, the word order in some of the translated sentences was subjectively better when the soft boundary penalty was applied. 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.

F. Extractive Speech Summarization

Extractive speech summarization algorithms [30], [2] 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.

Segmentations currently used or proposed for extractive summarization include words, phrases, sentences, or speaker turns [30]. 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. Each segment was labeled for inclusion in the summary if more than 50% of a segment was found in the human summary. Inclusion vs. exclusion was predicted automatically using a Bayesian network classifier which used only acoustic and structural features for summarization. Using the standard ROUGE summarization score, the best results were obtained with intonational phrases, with ROUGE-1 and ROUGE-L score of 0.57 and 0.56, respectively.

Other experiments by researchers at UT Dallas have looked at whether tuning the sentence segmentation threshold for the summarization application could lead to improved performance. In this case, experiments were on the ICSI meeting corpus, and the classifier used textual features only and the maximal marginal relevance for extractive summarization. They used an HMM for sentence segmentation, and varied the decision threshold from the segmentation system, and used different units for the subsequent

summarization module. The results showed that the performance was stable over a large range of sentence segmentation thresholds, with a ROUGE-1 score of roughly 0.63 for threshold ranging from 0.4 to 0.9.

V. SUMMARY

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 formating 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.

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 suggests the use of a low threshold (more hypothesized boundaries) with confidences associated with the boundaries, so that different downstream modules can use their own threshold. When alternative word hypotheses are represented with N-best lists, typically with fixed points at sentence boundaries, then shorter sentences can also mean more variation in the word hypotheses.

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. However, effective automatic identification of speech repairs and their structure is important for many speech processing applications like parsing [19] and machine translation.

Finally, it is important to remember that there is information in speech beyond what is in text, and it is a mistake to consider speech as just an impoverished alternative to text. There is certainly a benefit from leveraging segmentation to make speech look more like text due to the fact that language processing systems tend to be trained on text. However, there are cues to speaker intent and information salience that are also there to be mined in future applications.

REFERENCES

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

[2] J. Mrozinski, E. Whittaker, P. Chatain, and S. Furui, "Automatic sentence segmentation of speech for automatic summarization," in *Proc. Inter. Conf. Acoustics, Speech, and Signal Processing*, 2006, pp. 981–984.

[3] 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.

[4] M. Gregory, M. Johnson, and E. Charniak, "Sentence-internal prosody does not help parsing the way punctuation does," in *Proc. HLT/NAACL Conference*, 2004, pp. 81–88.

[5] 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.

[6] 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.

[7] 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.

[8] S. Chen and P. Gopalakrishnam, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Workshop*, 1998, pp. 127–132.

[9] 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.

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

[11] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

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

[13] B. Roark, Y. Liu, M. P. H. 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.

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

[15] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, 1998, pp. 3189–3192.

[16] E. Shriberg, S. Cuendet, B. Favre, J. Fung, and D. Hakkani-Tur, "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.

[17] 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.

[18] D. Hillard *et al.*, "Impact of automatic comma prediction on pos/name tagging of speech," in *Proc. IEEE/ACL Workshop Spoken Language Technology*, 2006, pp. 58–61.

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

[20] 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.

[21] B. Roark, M. P. Harper, E. Charniak, B. Dorr, M. Johnson, J. Kahn, Y. Liu, M. Ostendorf, J. Hale, A. Krasnyanskaya, M. Lease, I. Shafran, M. Snover, R. Stewart, and L.Yung, "SParseval: Evaluation metrics for parsing speech," in *Proc. Language Resource and Evaluation Conference*, 2006.

[22] 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.

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

[24] A. Stolcke *et al.*, "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1729–1744, 2006.

[25] S. Maskey and J. Hirschberg, "Soundbite detection in broadcast news domain," in *Proc. Interspeech*, 2006, pp. 1543–1546.

[26] F. Liu and Y. Liu, "Look who is talking: Soundbite speaker name recognition in broadcast news speech," in *Proc. HLT/NAACL Conference*, 2007, pp. 101–104.

[27] 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.

[28] 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.

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. Annual Meeting Assoc. Comp. Ling.*, 2002, pp. 311–318.

[30] 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.