Transcribing Human-Directed Speech for Spoken Language Processing

Mari Ostendorf

Department of Electrical Engineering, University of Washington, Seattle, Washington

mo@ee.washington.edu

Abstract

As storage costs drop and bandwidth increases, there has been a rapid growth of spoken information available via the web or in online archives, raising problems of document retrieval, information extraction, summarization and translation for spoken language. While large is a long tradition of research in these technologies for text, new challenges arise when moving from written to spoken language. In this talk, we look at differences between speech and text, and how we can leverage the information in the speech signal beyond the words to provide structural information in a rich, automatically generated transcript that better serves language processing applications. In particular, we look at three interrelated types of structure (orthographic, prosodic, and syntactic), methods for automatic detection, the benefit of optimizing rich transcription for the target language processing task, and the impact of this structural information in tasks such as information extraction, translation, and summarization.

Index Terms: prosody, parsing, rich speech transcription

1. Introduction

Large collections of speech recordings are increasingly available via the Internet, including radio and TV broadcasts, legislative proceedings, debates, lectures, hearings, oral histories, and webcasts, among other types of communications. Companies also keep large archives of call center data, and even personal voicemail collections can become large. We will refer to these human-directed (vs. computer-directed) speech signals as spoken documents. Listening to speech requires much more time than skimming transcripts, so automatic spoken language processing is necessary to make any real use of this data. Dramatic improvements in automatic speech recognition (ASR) technology have resulted in improved quality of word transcripts, which facilitates indexing and retrieval. However, an unstructured sequence of words is not as rich as written text, and not generally what most language processing systems expect as input. For example, translation, information extraction and summarization systems all work with sentence units, though they may be determined automatically from punctuation.

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, 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, taken from a talk show. Human listeners can easily segment such spoken input, arriving at the subsequent formatted version. To do so they can draw on a combination of syntactic, semantic, acoustic, prosodic, pragmatic, and discourse knowledge, though we do not fully understand this process.

Unformatted Word Transcripts

we'll be sharing those comments coming up during the segment very good interested to see what the folks have to say out there first let's see what our bloggers have to say conservative writer commentator matt lewis of politics daily dot com we'll try to get you back

Formatted/Cleaned Word Transcripts

Long: We’ll be sharing those comments coming up during the segment.
Aqui (host): Very good. Interested to see what the folks have to say out there. First let’s see what our bloggers have to say. Conservative writer commentator Matt Lewis of politicsDailly.com. Matt, thanks for being there.
Lewis: Hello
Aqui: Hello. Next up gay rights activist and writer Bill Browning, editor-in-chief of the Bilerico Project of bilerico.com. Did we just lose you Bill? We’ll try to get you back.

Automatic systems can use acoustic-prosodic cues and recognized words to insert these orthographic (surface-form) indicators of structure, though not yet at a very high accuracy level. Having such markup, along with simple inverse text normalization transformations (e.g., of word strings into numbers) and truecasing, makes speech more like text and therefore makes it possible to use linguistic resources based on written text for training spoken language processing systems. Since annotated resources for written text are much more extensive than those for speech, and since written text is more easily harvested from the web, it is essential to be able to take advantage of text resources.

That said, it is important to acknowledge that spoken language differs from written language in ways other than the absence of explicit punctuation. Specific text and speech genres are characterized by differences in word choices and complexity of sentence structure, as is taken advantage of in text classification. For example, spoken language tends to include more word-based discourse markers, like “well,” “so,” “now,” etc. Conversational speech, in particular, is generated with “real-time” planning and thus contains disfluencies and filled pauses. It also tends to contain more frequent use of the pronouns “I” and “you” and more informal word choices than other genres. Further, speech differs from text in that intonational cues add to
the information in the words, providing indicators of emphasis or contrast, intent (e.g. question, statement, doubt, etc.), and discourse structure (topic, turn-taking). Language processing systems could take advantage of these cues, though only if departing from a strict application of text processing methods.

In addition to using orthographic cues to sentence structure and word type, text-based language processing systems often use a parse of the sentence as an intermediate processing step. For example, parses are useful in information extraction for complex name phrases, and syntactic structure may be used for a pre-processing re-ordering step in translation. Summarization that involves redundancy removal typically also uses syntactic analysis. Applying these same systems to speech requires parsing spoken utterances. However, the problem of parsing speech is complicated by the different word choices, the potential for disfluencies, and the uncertainties associated with ASR errors. Parsers for speech need to be robust and matched to the spoken genre, which can involve explicit disfluency modeling.

In this paper, we describe approaches for annotating surface-form and symbolic (hidden) prosodic structure in speech (Section 2), and argue for a parsed representation as an objective for speech transcription (Section 3). Examples showing the impact of leveraging different types of structural information in language processing applications lead to the finding that rich transcription should be optimized for the language processing task and not transcription per se (Section 4). We note that most application studies involve annotation of surface form structure. Practical considerations of learning computational models for text/speech domain transfer as well as for hidden prosodic structure point to semi-supervised learning as an important tool for generalizing to new genres and leveraging text-based training resources (Section 5). Finally, we conclude by raising open questions and highlighting important directions for future research (Section 6).

2. Rich Transcription of Speech

Two types of structural information are useful for speech transcript markup: explicit surface-form orthographic markup and hidden structure such as prosodic events and syntax. In either case, automatic prediction of markup benefits from both word-based and acoustic cues. Word-based 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 constituent boundary (“the problem”), and for representing words that are likely to start a new sentence (such as “so,” in conversational speech). Acoustic features include spectral cues to speaker change and overlap, speaker identity, laughter, etc. Acoustic-prosodic features reflect information about duration, pause, intonational and energy contours, which are often extracted from automatic alignments of word and phone transcriptions with the speech signal. Typically the features are normalized to account for speaker, channel, and contextual effects, such as speaking rate and segmental context.

2.1. Surface-form structure

As illustrated in the example, one approach to enriching the word transcript provided by an ASR system is to predict speaker turns, sentence-final and other punctuation, inverse-normalize for verbalized forms, and predict case for languages such as English. This type of information will be referred to as surface-form structure. An advantage of surface-form structure is that prediction models can be learned from text-based resources, assuming reliance on lexical cues alone. To the extent that transcribed speech is available, acoustic cues can be leveraged through model combination or factoring the model to separate language and acoustic cues. An example of a factored model is the HMM-like hidden event model [1], which has been used extensively for tasks such as sentence and topic boundary prediction. The language model can be trained on text, and the prosodic “observation” model (decision tree, neural network, SVM, boosting tree, etc.) is trained on a subset of the transcripts that are aligned to audio files.

The hidden event model is an example of a tagging approach, where each word is associated with a hidden tag (e.g., +/- sentence boundary, comma, null) and the model predicts a tag for each word. Other tagging models (e.g. maximum entropy or conditional random fields) can also be used successfully, as explored in [2]. Tagging models focus on local cues, which is effective for many types of structure, but some problems (e.g. sentence-final punctuation or topic segmentation) benefit from features extracted given the beginning and end points of the constituent. An alternative to the tagging approach is to model whole constituents, which is also useful when maximum and/or minimum length constraints are needed. One example of a whole constituent model for sentence segmentation uses a log-linear model to combine word-based and prosodic features [3]. Speaker diarization systems also represent whole constituents [4]. The challenge of modeling the constituent is that the search space is much larger than when searching for sequential boundary events based on local cues, since all possible previous segment boundaries up to the maximum must be considered. Typically the set of possible candidates is restricted using a tagging-based first pass combined with simple heuristics.

2.2. Hidden structure

Two broad types of hidden structure are useful for spoken language processing: i) within-sentence prosodic events, including prosodic prominence or emphasis, prosodic phrase boundaries and disfluency interruption points; and ii) discourse-level information such as topic segmentation and speech act (or, dialog act) labels for utterances. As for surface-form structure, detection benefits from access to both word-based and acoustic cues.

Since the invention of prosodic annotation standards such as Tones and Break Indices (ToBI) [5] and the availability of annotated corpora, there has been much interest in building classifiers to automatically recognize symbolic prosody classes. Recognition of the different types of prosodic events have been explored in many prior studies, typically combining acoustic and lexical/syntactic features; e.g. [6, 7, 8, 2, 9, 10, 11]. Many of the same event detection models used for surface-form structure are also applied for prosody recognition, with boundary events being associated with the words that they follow and prominence with the accented word. Prosodic phrase structure is useful for resolving syntactic ambiguities, though there is not a deterministic mapping between the two sets of constituents. Prosodic phrases are potentially also useful as units in summarization or for constraining reordering. Prominence has been show to be useful in topic recognition [12] and in extracting key words and phrases for summarization [13]. Disfluency detection is useful in parsing as discussed in the next section.

There is a large literature on topic segmentation for text, mostly aimed at representing lexical coherence within a topic.
In speech, topic segmentation can also take advantage of acoustic cues, such as pitch range changes and speaker changes. While the pitch range changes may be a local phenomenon that could be captured with a tagging model, lexical coherence requires knowledge of the hypothesized start and end of the topic. Depending on the emphasis of the work in terms of features, very different types of statistical models have been proposed, e.g. [14, 15, 16, 17, 18, 19]. Because a topic segmentation creates coherent subsequences of words, the topic boundaries can be useful in passage retrieval and for specifying regions for language model adaptation.

Speech acts are categories of utterance function in a conversation, such as statement, question, backchannel, apology, greetings, etc. Recognizing dialog acts can be useful in summarization [20], as well as in determining speaker role in a conversation and in analyzing human-human interactions for communication effectiveness and for labeling discussion segments in terms of “meeting act” types such as brainstorming, review, argument, etc. [21]. Recognizing dialog acts typically involves both segmentation and type classification. (Since spoken “sentences” are not always grammatically complete, dialog act segmentation often stands in for sentence segmentation.) Segmentation can be a first step in a 2-step process to simplify the implementation, assuming that the dialog act boundary can be easily detected. Alternatively, the segmentation and classification can be done jointly, as in word recognition. Recognition of dialog acts has been investigated in several studies, with an important early effort being [22]; examples of later studies include [23, 24, 25, 26, 27].

3. Parsing Speech

As discussed previously, many language processing tasks require parsing, which is complicated for spoken language because of the lack of explicit punctuation and word uncertainty, as well as the presence of disfluencies.

Early work in parsing speech addressed the problem of disfluencies assuming known sentence boundaries. Initial studies used rule-based methods [28, 29]; a statistical parsing approach was later explored in [30] using a two-stage approach of edit detection and then parsing. An integrated approach that used a TAG-based model for edits was introduced in [31], and the role of prosodic cues was explored in [32]. These studies provided important baselines, but did not address the real problem where the sentence segmentation is unknown and word transcriptions may have errors.

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, automatic sentence segmentation provides better matching of training and testing conditions and can improve accuracy. Studies assessing the effect of sentence segmentation quality on parsing showed that automatic (vs. reference) sentence segmentation, since ASR accuracy has a big impact on sentence segmentation.

One can argue that the SParseval measure is a good objective for ASR in language processing applications, because a parsed transcript is the desired input for many language processing applications. However, even for those systems that do not leverage syntax, the SParseval measure may be useful because it effectively puts more weight on “important” words, unlike the standard word error rate (WER) measure, which assumes that all words have equal importance. Dependency-based SParseval computes an F-score indicating the degree to which triples match between a word-pare hypothesis and a reference parse, a score which is effectively weighting the headwords as more important. For example, the main verb is often the most important word in terms of its role as a headword in multiple dependencies – appropriate, given that main verb errors can cause substantial problems for language processing. Conversely, SParseval treats determiners (which do not appear as headwords) as least important. In analysis of speech recognition output and Mandarin-to-English translation performance as measured by human-targeted translation edit rate (HTER), SParseval is found to be significantly better correlated with HTER than WER, particularly for low-WER sentences [38]. Notably, the correlation was not high when using a Mandarin parser trained on newswire for talkshow data, but after adapting the parser to the talkshow genre (using semi-supervised learning, as discussed below), we found that optimizing the recognizer for a SParseval objective did in fact lead to improvements in MT.

4. Integration and Impact on Language Processing

Many studies have shown that automatic sentence segmentation has a significant impact on language processing. For example, in the previous section, we described studies showing that sentence boundary detection has a significant impact on parser performance. Another study [39] confirmed these observations for English information extraction (IE) on speech, using the NYU IE system [40] and a subset of TDT4 English broadcast news corpus. They found that optimizing sentence
and comma prediction thresholds for IE performance is more effective than optimizing these thresholds separately for punctuation prediction accuracy: improvements in the ACE scores are from 15.6 to 18.4 for relations, and 47.0 to 48.2 for entities. Error analysis showed that punctuation errors can result in merged noun phrases or split entities. The best case performance was obtained by jointly optimizing comma and sentence boundary thresholds but allowing the thresholds to vary in detecting entities vs. relations. Automatic comma prediction was also shown to help name tagging in Mandarin broadcast news [41].

Different sentence segmentation algorithms have also been evaluated on large vocabulary Arabic-to-English and Chinese-to-English broadcast news translation tasks using the phrase-based MT system of RWTH [42, 43]. MT performance improves by combining word-based (tagging) and whole constituent methods. For broadcast news tasks, the approach increased BLEU from 18.1 for fixed-length segments to 21.2 for the MT-optimized sentence predictions. The best MT result was achieved by using a phrase coverage feature that provides an indicator of whether an added boundary will eliminate the potential to use a frequent phrase. The sentence boundary precision is actually reduced significantly 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.

Experiments on tuning the sentence segmentation thresholds for extractive summarization have reported mixed results. In experiments on the ICSI meeting corpus with HMM-based sentence segmentation and sentence selection using a maximal marginal relevance with textual features only, results showed that the best sentence detection threshold for summarization was similar to that for optimizing sentence detection directly [44]. However, in experiments on English broadcast news and different types of units (intonal phrases, pause-based chunking and sentences) as alternatives for segmentation in summarization with an oracle sentence selection mechanism, the best results were obtained with intonational phrases [45], which are shorter than sentences on average. The difference may be a consequence of the specific text features used in the fully automatic result.

A common feature of all of these results is the need to optimize the segmenter for the target language processing task. In many cases, though not all, better results were obtained with shorter segments. The fact that different applications benefit from different thresholds argue for an approach of providing soft decisions associated with confidences. It may also be a consequence of the need for a hierarchical segmentation structure.

Note that it is primarily surface-form cues that have been utilized in spoken document processing (vs. speech understanding). The hidden structural categories described previously are fundamentally speech-based, and therefore data-driven learning requires some annotated speech corpora. However, hand-annotation is expensive, and the studies reported on hidden structure recognition tend to be restricted to only a few relatively small corpora which may not generalize to larger tasks, particularly because of the small number of speakers represented. Alternatively, some studies have used acoustic-prosodic features directly, rather than symbolic events, e.g. [46, 47, 48]. The fact that there is little data annotated with prosodic cues, in dialog acts and other aspects of discourse structure has been a limiting factor in the use of this type of structure in language processing. However, advances in semi-supervised learning can have an impact on this problem.

5. Semi-supervised learning
Semi-supervised learning, which broadly involves leveraging a combination of labeled and unlabeled data, is useful in cases where only a small amount of target data is labeled, which is valuable for reducing the need for hand annotation. In addition, many of the same ideas can be applied to the problem of domain transfer or unsupervised domain adaptation, where there is annotated data for one or more domains (defined in terms of application or genre) that might be leveraged in learning a model for a different domain. Analyses of prosodic features in meeting recordings vs. broadcast news in English show that F0, duration, and energy features show remarkable similarity across styles [49], which offers hope for the domain transfer approach. Recalling our earlier observations about the need to rely on written text sources for spoken language modeling provides another argument for exploring semi-supervised learning in the context of domain adaptation.

There are many different approaches to semi-supervised learning, including variations on the idea of automatically labeling the unlabeled data, graph-based learning, and methods accounting for feature space differences. The research on semi-supervised adaptation and domain adaptation in general are extensive; two surveys on the respective topics are [50, 51].

In speech structure annotation, the most widely used methods are based on automatic labeling (also called bootstrapping), specifically self-training and co-training, which iteratively introduce labels on unlabeled data. The Expectation-Maximization (EM) algorithm [52] can be used as a form of self-training, where an initial model is trained on labeled data, unlabeled samples are assigned to different classes with some probability according to the previous model, and a new model is estimated from these weighted counts. Co-training [53] is a popular semi-supervised learning algorithm, which exploits two conditionally independent views of the same data assuming that each subset is sufficient for classifier learning. Two classifiers are trained on the labeled data using the different views, and then each classifier is used to select high confidence data points to add to the training set of the other classifier. The labeling process is iterated, increasing the sets of labeled data. The EM and co-training ideas have also been combined in co-EM [54, 55]. The multi-view approach is a good match to problems in modeling spoken language structure, since there is a natural split into word-based and prosodic features. Experiments have investigated the use of different variations of self- and co-training in sentence and dialog act segmentation [56, 57] and in dialog act tagging [58, 59]. Co-training and self-training have also been used by many researchers in parsing. Of particular relevance to spoken language processing is its use in genre adaptation, as in [60, 61].

Differences that have been observed in prosodic structure associated with speaking styles can in some cases be described by a change in class posteriors. For example, careful radio news includes more frequent pitch accents than conversational speech, and more sentence-internal intonational phrase boundaries. When features are roughly similar across domains, the covariate shift approach (e.g., [62]) provides a simple resampling solution to the domain adaptation problem.

Unsupervised learning might also be useful for some tasks; indeed, good results were presented for prominence detection [63]. However, a challenge in unsupervised learning is that the acoustic cues to the types of hidden structural events that we
have proposed are overlapping. For example, phrase-final syllables and prominent words are both lengthened, albeit with different patterns. Pauses may occur at fluent phrase boundaries, as well as in disfluent regions, and occasionally to set off an emphasized word. Both sentence and topic boundary events and pitch accent types are characterized by intonation and energy cues. The details of the realization and the interaction of different cues make it possible to discriminate the different classes, but simply clustering vectors of these raw features may not lead to useful categories. Another complication is that a commonly used acoustic feature is word duration, but word duration at a coarse level is simply indicative of the content vs. function word distinction. Unfortunately, a prosodic classifier based on such a simple heuristic is of limited use for many applications. Thus, taking more extensive advantage of unsupervised learning will likely require further research on feature extraction and vector similarity functions, which would require some labeled data.

6. Discussion

In summary, we have brought together a series of studies – on parsing, information extraction, translation, and summarization – that make the case for moving beyond transcription of speech as a “string of words,” showing that language processing on human-directed speech benefits from including multiple types of segmentation (topic, speaker, sentence). Further, we argue that parsing should be integrated with recognition and that a parse score should replace word error rate as an objective for speech transcription.

Experiments consistently show a benefit from optimizing segmentation for the application. This raises the question of how to choose an operating point when multiple types of processing are involved. One solution is simply to provide boundary confidence posteriors (or word prominence posteriors, etc.) at each hypothesized word boundary and allow different thresholds for different tasks. Another alternative is to represent a hierarchy, including prosodic phrases, and optimize for application performance over the choice of unit rather than the detection threshold for a particular unit.

We observed that most large-scale spoken document processing leverages only orthographic structure; prosodic events have mainly been explored in small data sets. If prosodic cues are important to human listeners, as we know from linguistic studies and from research on speech synthesis, why have they not been more universally leveraged in spoken document processing applications? One explanation is the annotation problem. For synthesis studies, annotation of data from a few target speakers is enough, but for recognition applications, it is critical to have training data from a large number of speakers. Advances in semi-supervised learning can have impact here, both for recognition and for synthesis as voice transformation takes on increasing importance. Further research in the acoustic correlates of prosodic events will certainly also benefit this effort.

7. Acknowledgments

The author would like to thank collaborators at ICSI, SRI, University of Maryland, RWTH, Columbia University, the University of Texas at Dallas, NYU, CUNY and Brown, who have made it possible to explore the impact of rich transcription of speech in a wide variety of applications. In addition, several University of Washington graduate students have been responsible for many of the results and insights summarized here, including C. Boulis, S. Hahn, D. Hillard, J. G. Kahn, and A. Margolis.

8. References


