STRUCTURAL METADATA RESEARCH IN THE EARS PROGRAM

Yang Liu[1,5] *Elizabeth Shriberg*[1,2] *Andreas Stolcke*[1,2]
Barbara Peskin[1] *Jeremy Ang*[1] *Dustin Hillard*[3]
Mari Ostendorf[3] *Marcus Tomalin*[4] *Phil Woodland*[4] *Mary Harper*[5]

[1]International Computer Science Institute, USA [2]SRI International, USA
[3]University of Washington, USA [4]Cambridge University, UK [5]Purdue University, USA
{yangl,ees,stolcke,barbara}@icsi.berkeley.edu

ABSTRACT

Both human and automatic processing of speech require recognition of more than just words. In this paper we provide a brief overview of research on structural metadata extraction in the DARPA EARS rich transcription program. Tasks include detection of sentence boundaries, filler words, and disfluencies. Modeling approaches combine lexical, prosodic, and syntactic information, using various modeling techniques for knowledge source integration. The performance of these methods is evaluated by task, by data source (broadcast news versus spontaneous telephone conversations) and by whether transcriptions come from humans or from an (errorful) automatic speech recognizer. A representative sample of results shows that combining multiple knowledge sources (words, prosody, syntactic information) is helpful, that prosody is more helpful for news speech than for conversational speech, that word errors significantly impact performance, and that discriminative models generally provide benefit over maximum likelihood models. Important remaining issues, both technical and programmatic, are also discussed.

1. INTRODUCTION

Although speech recognition technology has improved significantly in recent decades, current speech systems still output simply a stream of words. This unannotated word stream does not include useful information about punctuation and disfluencies. Such structural information is important for speech transcripts to be human readable [1]. It is also crucial for effective use of subsequent natural language processing techniques, which are typically based on the assumption of fluent, punctuated, and formatted input. Recovering structural information in speech has thus become the goal of a growing number of studies in computational speech processing, e.g., [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. The metadata extraction (MDE) research effort within the DARPA EARS program [12] aims to enrich speech recognition output by adding automatically tagged information on the location of sentence boundaries, speech disfluencies, and other important phenomena. In this paper, we focus on automatically detecting structural information in the word stream (the so-called "structural MDE" portion of the EARS program); other MDE efforts on speaker diarization are overviewed in [13].

The rest of this paper is organized as follows. We describe the structural MDE tasks, performance measurement, and corpora for the EARS program in Section 2. Section 3 introduces general approaches used for structural MDE. Results are presented in Section 4. A summary and discussion of open issues appear in Section 5.

2. MDE TASKS

2.1. Task Description

Several structural events are annotated in the EARS program. These include: sentence-like units (SUs), edit disfluencies, and filler words (see [14] for annotation guidelines). Corresponding to these events, the Rich Transcription structural MDE framework includes four tasks.

- SU detection aims to find the end point of an SU. The detection of subtype (statement, backchannel, question, or incomplete) for each SU is also required.
- Edit word detection aims to find all words within the reparandum region of an edit disfluency. These are the words that will be removed to obtain cleaned-up transcripts.
- Filler word detection aims to identify words used as filled pauses (e.g., *uh, um*), discourse markers (e.g., *you know, like, so*), and explicit editing terms (e.g., *I mean*).
- Interruption point (IP) detection aims to find the interword location at which point fluent speech becomes disfluent. This includes the interruption point inside an edit disfluency and the starting point of a filler word string.

The following example shows a transcript with metadata marked: './' for statement SU boundaries, '< >' for fillers, '[ ]' for edit words, and '*' for IPs inside edit disfluencies.

```
and  < uh > < you know > wash your clothes
wherever you are ./ and [ you ] * you really
get used to the outdoors ./
```

2.2. Performance Measures

Each task is evaluated separately. The NIST scoring tools created for these tasks first align the reference and hypothesis words to minimize the word error rate. After alignment, the hypothesized structural events are mapped to the reference events using the word alignment information, and then unmatched structural events are counted. For edit and filler word detection, the error rate is the average number of misclassified reference tokens per reference edit or filler word token. For SU and IP detection, the error rate is the average number of misclassified boundaries per reference SU or

In *Proc. ICASSP 2005, Volume V, pp. 957–960.*

IP. The error rate in the NIST metric can be greater than 100% due to insertions. A detailed description of the scoring tool is provided at http://www.nist.gov/speech/tests/rt/rt2004/fall/.

Standard tests for the significance of differences between systems have only recently been introduced, with NIST reporting results with the Wilcoxon signed rank test for speaker-level average score differences. While a range of techniques is used in word error rate scoring for speech recognition, so far only the speaker-level test and a pause unit matched pair test have proved useful for metadata scoring [15].

A limitation of the standard MDE scoring methods is that they examine only one operating point out of a range of possibilities, and there may be different false-alarm/missed detection tradeoffs that make sense for downstream language processing applications. Further, most researchers prefer soft decisions, i.e., decisions with confidence scores that can be used as weights with other knowledge sources. If confidence scores are given at each interword boundary, systems can pick the best operating point for their application. Mechanisms for evaluating the performance range using a decision-error tradeoff (DET) curve or receiver operating characteristic (ROC) curve are proposed in [15] and [16].

2.3. MDE Corpora

Conversational telephone speech (CTS) and broadcast news (BN) are used for the structural event detection tasks in EARS. CTS and BN are very different genres. They differ for example in the average SU length and frequency of disfluencies. Speech in BN has fewer disfluencies, sentences are longer and more grammatical, and the speakers are mostly professionals reading teleprompted text. Speech in CTS is more casual and conversational, containing many backchannels, filler words, and edit disfluencies. For each corpus, two different types of transcriptions are used: human-generated transcription (REF) and speech-to-text recognition output (STT). Using the reference transcriptions provides the best-case scenario for the evaluation of a structural event detection algorithm because there are no word errors in the transcriptions.

Table 1 shows the distribution of different structural events in the two corpora (measured by the percentage of the interword boundaries that are labeled with the events), along with the size of the training and testing sets in the most recent Rich Transcription evaluation (RT-04), and the word error rate (WER) on the test set obtained from the best speech recognition output in the RT-04 evaluation (from a multiple system combination). The statistics for the development sets are similar to the eval test sets. Additionally, there is training data annotated with an earlier version of the annotation guideline, but that data is not always used due to the changes in the annotation guidelines.

	CTS	BN
Training set (number of words)	484K	182K
Test set (number of words)	35K	45K
STT WER (%)	14.9	11.7
SU percentage	13.6	8.1
Edit word percentage	7.4	1.8
Filler word percentage	6.8	1.8

Table 1. Information on the CTS and BN corpora used in the most recent RT-04 evaluation, including the data set sizes, the recognition WER on the test set, and the percentage of the different types of structural events in the training set.

3. MDE SYSTEM

The MDE tasks can be seen as classification tasks that determine whether an interword boundary is an event boundary (e.g., SU or IP) or whether a word belongs to an event of interest. In this section, we describe system approaches used for these tasks and briefly summarize previous work.

3.1. Knowledge Sources

Most of the MDE systems use both textual and prosodic information. Typically, at each interword boundary, prosodic features are extracted to reflect pause length, duration of words and phones, pitch contours, and energy contours. These prosodic features are modeled by a classifier (e.g., a CART decision tree), which generates a posterior probability of an event given the feature set associated with a boundary. Textual cues are captured by contextual information of words or their corresponding classes or higher-level syntactic information. For example, an N-gram language model (LM) can be used to model the joint probability $P(W, E)$ of the word and the event sequence. A transformation-based learning (TBL) classifier is used in [5, 7] to capture textual knowledge for disfluency detection.

3.2. Frameworks for Combining Knowledge Sources

An HMM is commonly used to combine the two knowledge sources (prosodic and textual) [17, 18]. In this framework, the transition probabilities are modeled generally by a hidden event N-gram LM. Task-specific LMs are often used to model the token sequences associated with each MDE task. Different LMs (word and class based) have also been interpolated [17, 19]. The observation (F) probability $P(F|E)$ is obtained from the prosody model that generates $P(E|F)$. Various decoding techniques have been explored including 1-best Viterbi decoding, posterior decoding, and forward-backward decoding [8, 18].

Recently, studies using maximum entropy (Maxent) and conditional random fields (CRF) have been conducted, in an attempt to address the weakness of the generative HMM approach [17, 20]. These approaches directly estimate the posterior probability of an event given observations and better match the performance metrics. Additionally, they provide more freedom for incorporating various knowledge sources, especially overlapping features.

CRF and Maxent differ from an HMM with respect to the training objective function (joint versus conditional likelihood) and their handling of overlapping word-related features. HMM training does not maximize the posterior probabilities of the correct label; while the CRF and Maxent models directly estimate posterior boundary label probabilities. The underlying N-gram sequence model of an HMM does not cope well with multiple representations of the word sequence (e.g., words, part of speech); however, the CRF and Maxent models support simultaneous correlated features. The CRF and HMM differ from the Maxent method with respect to their ability to model sequence information. The Maxent model only makes decision locally.

3.3. Related Work

Much research has been devoted to automatically detecting structural information from text or speech prior to the EARS program. Past work has shown that both textual and prosodic cues provide important information for the detection of sentence boundaries and disfluencies. Most of these experiments were conducted on human transcriptions, many focused on only one corpus or task, and some prior studies on disfluency detection relied on the assumption that sentence boundary information is available. The MDE effort in the EARS program aims to explore these tasks more extensively, using different corpora and different transcriptions, across different tasks. Most important, the main goal is to rely on speech only, that is, using recognition output and without assuming the availability of any structural information.

4. SYSTEM PERFORMANCE

Due to space limitations, we focus in the remainder of the paper on SU/SU-subtype detection and edit detection. We omit filler word detection, for which reasonable results can be achieved with simple text-based classifiers. See [17] for more discussion about the filler word detection task.

4.1. SU/SU-subtype Detection

The most widely used approach for this task is an HMM combining an N-gram LM and a CART decision tree prosody model. Since SU boundary events are much rarer than the nonevents, sampled training sets are generally used to train a decision tree to make it more sensitive to the inherent properties of the events [18]. Liu *et al.* [21] applied bagging and various sampling methods to obtain more reliable posterior probability estimations for the prosody model. Various textual features (class-based LMs and LMs trained using auxiliary annotated data) are used, in addition to the word-based hidden event LM that is trained from the LDC annotated training data [17, 18]. In [17], Maxent and CRF models were investigated, both of which use features from N-grams of words and classes, the binned posterior probabilities from the prosody model and from the LM trained using extra text corpora. Combinations of these approaches are also used to obtain the SU boundary hypotheses. After SU boundaries are detected, a second step is used to determine the subtype of the SUs using a Maxent classifier [17].

		SU boundary error	SU total error
CTS	REF	26.21	36.80
	STT	39.18	49.24
BN	REF	47.15	49.71
	STT	59.73	61.95

Table 2. Results (%) of SU detection for BN and CTS, on REF and STT conditions. Subtype substitution errors are ignored in the "boundary error" and included in the "total error."

Table 2 shows SU detection results reported in [17], using the majority vote of the HMM, Maxent, and CRF approaches on CTS, and the linear posterior probability interpolation of the HMM and Maxent on BN. The SU error rate is higher on BN, suggesting that it is a harder task than CTS. This is partly because BN sentences are more complex, and the sparse data problem is more severe for BN; whereas, in CTS pronouns and backchannels are frequent and are good predictors for SU boundary detection. System performance degrades significantly using the recognition output rather than reference transcriptions, as indicated by both detection errors in Table 2 and the DET curves in Figure 1. The difference between the SU boundary detection error and the total error (i.e., SU substitution error) is smaller on BN than on CTS since almost all SUs are statements on BN.

Detailed analysis [16] has shown that adding textual information, building a more robust prosody model, using conditional modeling approaches (Maxent and CRF), and system combination all yield performance gains. Additionally, textual information is affected more by word errors in the recognition output condition than is the prosody model. Experiments in [17] have also shown that for BN obtaining speaker information from the speaker diarization results generally outperforms using simple speaker clustering as implemented for adaptation in speech recognition.

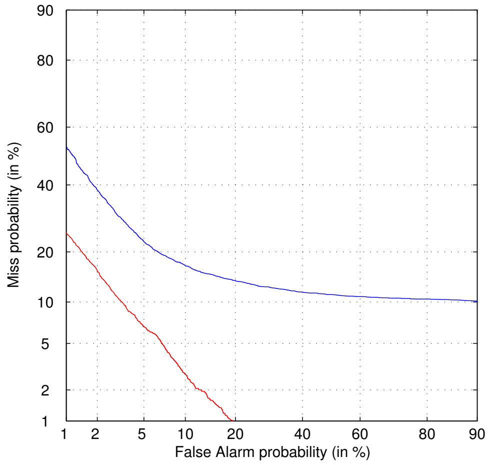

Fig. 1. DET curve for SU detection based on confidence predictions for the CTS reference transcript (lower curve) and STT output (upper curve).

To address the issue of higher SU detection errors on recognition output, Hillard *et al.* [22] extend the SU detection system of [17] to detect SU boundaries on multiple recognition hypotheses. The detected boundaries for each hypothesis are then combined using confusion networks and produce a small reduction in error for the CTS SU boundary detection task.

4.2. Edit Word Detection

Liu *et al.* [17] investigated detecting edit words and edit IPs using three modeling approaches. First, an HMM is used to combine the hidden event LM and a prosody model for IP detection. Heuristic rules are then used to find the onset of the reparandum. A separate repetition detector is used to detect repeated words. Second, a Maxent classifier is used to find the IP. Then, like the HMM, a rule-based approach is used to find the extent of the edit words. Third, a CRF model is implemented that detects the edit region and IP jointly. In this model, each word has an associated tag, representing the position of the word in the edit, such as at the beginning, inside, and outside of an edit. The Maxent and CRF approaches have shown to generally outperform the HMM for edit word detection.

		Edit word error
CTS	REF	50.07
	STT	80.41
BN	REF	43.00
	STT	89.86

Table 3. Results (%) for edit word detection for BN and CTS on REF and STT conditions.

Table 3 shows results from [17] for edit word detection, which used the CRF approach for CTS and the Maxent model for BN. The system degrades even more in the STT condition than for the

SU task, in part because word fragment information (an important indicator for edit disfluencies) is unavailable in the STT condition. In addition, it may be that edit detection relies more on word cues (e.g., repeats) than SU detection. Lease *et al.* [23] used a Tree Adjoining Grammar for edit word detection, and achieved better results than those shown in Table 3, suggesting that better modeling of the correspondences between words in the reparandum and corrections in disfluencies may be needed for MDE.

5. SUMMARY AND OPEN ISSUES

Finding structural information is important for improving transcript readability and aiding downstream language processing modules. We have provided a brief overview of research on structural metadata extraction in the DARPA EARS program. Approaches to automatic detection generally combine lexical and prosodic information, using various modeling techniques for knowledge source integration. The performance of these methods is evaluated by MDE task, by data source, and by whether input transcriptions to the system come from humans or from an (errorful) automatic speech recognizer. We have shown representative results for the SU and edit tasks. Results show that combining multiple knowledge sources (words, prosody, syntactic information) is helpful, that prosody is more helpful for BN than for CTS, that word errors significantly impact performance (but differentially for different tasks and corpora), and that discriminative models generally provide benefit over maximum likelihood models.

While great progress has been made in this area, which constitutes a new direction of research for DARPA, several open technical and programmatic issues remain. On the technical side, it is important to continue to search for better features; prosodic features in particular could be improved by using additional temporal context. Another issue is to develop better joint modeling for continuous and discrete features. We continue to look for features and models that are more robust to word recognition errors. Joint modeling of MDE events themselves is yet another technical focus area. Finally, it is important to learn to make use of partially-labeled or unlabeled training data.

On the programmatic side, one issue is how to achieve better interannotator agreement, and whether disagreement should be accounted for during scoring. A second issue is how to assess significance, since segmentation methods used for assessing word accuracy may not be appropriate for assessing structural phenomena. Third, should different tasks be scored separately, or integrated into a joint score? Additional questions concern extensions to new languages. Finally, researchers in the EARS community and beyond are beginning to look into the complex interaction between speech recognition, MDE, and downstream processing applications.

Acknowledgments

This research has been supported by DARPA under contract MDA972-02-C-0038 and MDA972-02-1-0013 and by NSF IRI-9619921. Distribution is unlimited. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA. Part of this work was carried out while the last author was on leave from Purdue University and at NSF.

6. REFERENCES

[1] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. of Eurospeech*, 2003.

[2] P. Heeman and J. Allen, "Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue," *Computational Linguistics*, vol. 25, pp. 527–571, 1999.

[3] J. Kim and P. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. of Eurospeech*, 2001, pp. 2757–2760.

[4] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*, 2000.

[5] M. Snover, B. Dorr, and R. Schwartz, "A lexically-driven algorithm for disfluency detection," in *Proc. of HLT/NAACL*, 2004.

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

[7] J. Kim, S. E. Schwarm, and M. Ostendorf, "Detecting structural metadata with decision trees and transformation-based learning," in *Proc. of HLT/NAACL*, 2004, pp. 137–144.

[8] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, pp. 127–154, 2000.

[9] D. Wang and S. S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *Proc. of ICASSP*, 2004.

[10] C. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech," *Journal of the Acoustical Society of America*, pp. 1603–1616, 1994.

[11] J. Bear, J. Dowding, and E. Shriberg, "Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog.," in *Proc. of ACL*, 1992, pp. 56–63.

[12] DARPA Information Processing Technology Office, "Effective, affordable, reusable speech-to-text (EARS)," http://www.darpa.mil/ipto/programs/ears/, 2003.

[13] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. of ICASSP*, 2005.

[14] S. Strassel, *Simple Metadata Annotation Specification V6.2*, Linguistic Data Consortium, http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf, 2004.

[15] D. Hillard and M. Ostendorf, "Scoring structural MDE: Towards more meaningful error rates," in *Proc. of EARS RT-04 Workshop*, November 2004.

[16] Y. Liu, *Structural Event Detection for Rich Transcription of Speech*, Ph.D. thesis, Purdue University, 2004.

[17] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, and M. Harper, "The ICSI/SRI/UW RT-04 structural metadata extraction system," in *Proc. of EARS RT-04 Workshop*, 2004.

[18] M. Tomalin and P. Woodland, "Advances in structural metadata for Eval04 at CUED," in *Proc. of EARS RT-04 Workshop*, 2004.

[19] M. Tomalin, S. Tranter, and P. Woodland, "SU detection for RT-03F at Cambridge University," http://www.nist.gov/speech/tests/rt/rt2003/fall/presentations/, 2003.

[20] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech," in *Proc. of EMNLP*, 2004.

[21] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Using machine learning to cope with imbalanced classes in natural speech: Evidence from sentence boundary and disfluency detection," in *Proc. of ICSLP*, 2004.

[22] D. Hillard, M. Ostendorf, A. Stolcke, Y. Liu, and E. Shriberg, "Improving automatic sentence boundary detection with confusion networks," in *Proc. of HLT/NAACL*, 2004, pp. 69–72.

[23] M. Lease, E. Charniak, and M. Johnson, "Parsing and its applications for conversational speech," in *Proc. of ICASSP*, 2005.