# Improving Automatic Sentence Boundary Detection with Confusion Networks

**D. Hillard**[†] **M. Ostendorf**[†]
University of Washington, EE[†]
{hillard,mo}@ee.washington.edu

**A. Stolcke**[*‡] **Y. Liu**[*] **E. Shriberg**[*‡]
ICSI[*] and SRI International[‡]
{stolcke,ees}@speech.sri.com
yangl@icsi.berkeley.edu

## Abstract

We extend existing methods for automatic sentence boundary detection by leveraging multiple recognizer hypotheses in order to provide robustness to speech recognition errors. For each hypothesized word sequence, an HMM is used to estimate the posterior probability of a sentence boundary at each word boundary. The hypotheses are combined using confusion networks to determine the overall most likely events. Experiments show improved detection of sentences for conversational telephone speech, though results are mixed for broadcast news.

## 1 Introduction

The output of most current automatic speech recognition systems is an unstructured sequence of words. Additional information such as sentence boundaries and speaker labels are useful to improve readability and can provide structure relevant to subsequent language processing, including parsing, topic segmentation and summarization. In this study, we focus on identifying sentence boundaries using word-based and prosodic cues, and in particular we develop a method that leverages additional information available from multiple recognizer hypotheses.

Multiple hypotheses are helpful because the single best recognizer output still has many errors even for state-of-the-art systems. For conversational telephone speech (CTS) word error rates can be from 20-30%, and for broadcast news (BN) word error rates are 10-15%. These errors limit the effectiveness of sentence boundary prediction, because they introduce incorrect words to the word stream. Sentence boundary detection error rates on a baseline system increased by 50% relative for CTS when moving from the reference to the automatic speech condition, while for BN error rates increased by about 20% relative (Liu et al., 2003). Including additional recognizer hypotheses allows for alternative word choices to inform sentence boundary prediction.

To integrate the information from different alternatives, we first predict sentence boundaries in each hypothesized word sequence, using an HMM structure that integrates prosodic features in a decision tree with hidden event language modeling. To facilitate merging predictions from multiple hypotheses, we represent each hypothesis as a confusion network, with confidences for sentence predictions from a baseline system. The final prediction is based on a combination of predictions from individual hypotheses, each weighted by the recognizer posterior for that hypothesis.

Our methods build on related work in sentence boundary detection and confusion networks, as described in Section 2, and a baseline system and task domain reviewed in Section 3. Our approach integrates prediction on multiple recognizer hypotheses using confusion networks, as outlined in Section 4. Experimental results are detailed in Section 5, and the main conclusions of this work are summarized in Section 6.

## 2 Related Work

### 2.1 Sentence Boundary Detection

Previous work on sentence boundary detection for automatically recognized words has focused on the prosodic features and words of the single best recognizer output (Shriberg et al., 2000). That system had an HMM structure that integrates hidden event language modeling with prosodic decision tree outputs (Breiman et al., 1984). The HMM states predicted at each word boundary consisted of either a sentence or non-sentence boundary classification, each of which received a confidence score. Improvements to the hidden event framework have included interpolation of multiple language models (Liu et al., 2003).

A related model has been used to investigate punctuation prediction for multiple hypotheses in a speech recognition system (Kim and Woodland, 2001). That system found improvement in punctuation prediction when rescoring using the classification tree prosodic feature model, but it also introduced a small increase in word error rate. More recent work has also implemented a similar model, but used prosodic features in a neural net instead of a decision tree (Srivastava and Kubala, 2003). A maximum entropy model that included pause information was used in (Huang and Zweig, 2002). Both finite-state models and neural nets have been investigated for prosodic and lexical feature combination in (Christensen et al., 2001).

### 2.2 Confusion Networks

Confusion networks are a compacted representation of word lattices that have strictly ordered word hypothesis slots (Mangu et al., 2000). The complexity of lattice representations is reduced to a simpler form that maintains all possible paths from the lattice (and more), but transforms the space to a series of slots which each have word hypotheses (and null arcs) derived from the lattice and associated posterior probabilities. Confusion networks may also be constructed from an N-best list, which is the case for these experiments. Confusion networks are used to optimize word error rate (WER) by selecting the word with the highest probability in each particular slot.

## 3 Tasks & Baseline

This work specifically detects boundaries of sentence-like units called SUs. An SU roughly corresponds to a sentence, except that SUs are for the most part defined as units that include only one independent main clause, and they may sometimes be incomplete as when a speaker is interrupted and does not complete their sentence. A more specific annotation guideline for SUs is available (Strassel, 2003), which we refer to as the “V5” standard. In this work, we focus only on detecting SUs and do not differentiate among the different types (e.g. statement, question, etc.) that were used for annotation. We work with a relatively new corpus and set of evaluation tools, which are described below.

### 3.1 Corpora

The system is evaluated for both conversational telephone speech (CTS) and broadcast news (BN), in both cases using training, development and test data annotated according to the V5 standard. The test data is that used in the DARPA Rich Transcription (RT) Fall 2003 evaluations; the development and evaluation test sets together comprise the Spring 2003 RT evaluation test sets.

For CTS, there are 40 hours of conversations available for training from the Switchboard corpus, and 3 hours (72 conversation sides) each of development and evaluation test data drawn from both the Switchboard and Fisher corpora. The development and evaluation set each have roughly 6000 SUs.

The BN data consists of a set of 20 hours of news shows for training, and 3 hours (6 shows) for testing. The development and evaluation test data contains 1.5 hours (3 shows) each for development and evaluation, each with roughly 1000 SUs. Test data comes from the month of February in 2001; training data is taken from a previous time period.

### 3.2 Baseline System

The automatic speech recognition systems used were updated versions of those used by SRI in the Spring 2003 RT evaluations (NIST, 2003), with a WER of 12.1% on BN data and 22.9% on CTS data. Both systems perform multiple recognition and adaptation passes, and eventually produce up to 2000-best hypotheses per waveform segment, which are then rescored with a number of knowledge sources, such as higher-order language models, pronunciation scores, and duration models (for CTS). For best results, the systems combine decoding output from multiple front ends, each producing a separate N-best list. All N-best lists for the same waveform segment are then combined into a single word confusion network (Mangu et al., 2000) from which the hypothesis with lowest expected word error is extracted. In our baseline SU system, the single best word stream thus obtained is then used as the basis for SU recognition.

Our baseline SU system builds on previous work on sentence boundary detection using lexical and prosodic features (Shriberg et al., 2000). The system takes as input alignments from either reference or recognized (1-best) words, and combines lexical and prosodic information using an HMM. Prosodic features include about 100 features reflecting pause, duration, F0, energy, and speaker change information. The prosody model is a decision tree classifier that generates the posterior probability of an SU boundary at each interword boundary given the prosodic features. Trees are trained from sampled training data in order to make the model sensitive to features of the minority SU class. Recent prosody model improvements include the use of bagging techniques in decision tree training to reduce the variability due to a single tree (Liu et al., 2003). Language model improvements include adding information from a POS-based model, a model using automatically-induced word classes, and a model trained on separate data.

### 3.3 Evaluation

Errors are measured by a slot error rate similar to the WER metric utilized by the speech recognition community, i.e. dividing the total number of inserted and deleted SUs by the total number of reference SUs. (There are no substitution errors because there is only one sentence class.) When recognition output is used, the words will generally not align perfectly with the reference transcription and hence the SU boundary predictions will require some alignment procedure to match to the reference location. Here, the alignment is based on the minimum word error alignment of the reference and hypothesized word strings, and the minimum SU error alignment if the WER is equal for multiple alignments. We report numbers computed with the su-eval scoring tool from NIST. SU error rates for the reference words condition of our baseline system are 49.04% for BN, and 30.13% for CTS, as reported at the NIST RT03F evaluation (Liu et al., 2003). Results for the automatic speech recognition condition are described in Section 5.

## 4 Using N-Best Sentence Hypotheses

The large increase in SU detection error rate in moving from reference to recognizer transcripts motivates an approach that reduces the mistakes introduced by word recognition errors. Although the best recognizer output is optimized to reduce word error rate, alternative hypotheses may together reinforce alternative (more accurate) SU predictions. The oracle WER for the confusion networks is much lower than for the single best hypothesis, in the range of 13-16% WER for the CTS test sets.

### 4.1 Feature Extraction and SU Detection

Prediction of SUs using multiple hypotheses requires prosodic feature extraction for each hypothesis, which in turn requires a forced alignment of each hypothesis. Thousands of hypotheses are output by the recognizer, but we prune to a smaller set to reduce the cost of running forced alignments and prosodic feature extraction. The recognizer outputs an N-best list of hypotheses and assigns a posterior probability to each hypothesis, which is normalized to sum to 1 over all hypotheses. We collect hypotheses from the N-best list for each acoustic segment up to 90% of the posterior mass (or to a maximum count of 1000).

Next, forced alignment and prosodic feature extraction are run for all segments in this pruned set of hypotheses. Statistics for prosodic feature normalization (such as speaker and turn F0 mean) are collected from the single best hypothesis. After obtaining the prosodic features, the HMM predicts sentence boundaries for each word sequence hypothesis independently. For each hypothesis, an SU prediction is made at all word boundaries, resulting in a posterior probability for SU and no_SU at each boundary. The same models are used as in the 1-best predictions – no parameters were re-optimized for the N-best framework. Given independent predictions for the individual hypotheses, we then build a system to incorporate the multiple predictions into a single hypothesis, as described next.

### 4.2 Combining Hypotheses

The prediction results for an individual hypothesis are represented in a confusion network that consists of a series of word slots, each followed by a slot with SU and no_SU, as shown in Figure 1 with hypothetical confidences for the between-word events. (This representation is a somewhat unusual form because the word slots have only a single hypothesis.) The words in the individual hypotheses have probability one, and each arc with an SU or no_SU token has a confidence (posterior probability) assigned from the HMM. The overall network has a score associated with its N-best hypothesis-level posterior probability, scaled by a weight corresponding to the goodness of the system that generated that hypothesis.

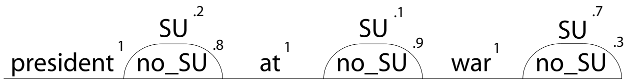

Figure 1: Confusion network for a single hypothesis.

The confusion networks for each hypothesis are then merged with the SRI Language Modeling Toolkit (Stolcke, 2002) to create a single confusion network for an overall hypothesis. This confusion network is derived from an alignment of the confusion networks of each individual hypothesis. The resulting network contains slots with the word hypotheses from the N-best list and slots with the combined SU/no_SU probability, as shown in Figure 2. The confidences assigned to each token in the new confusion network are a weighted linear combination of the probabilities from individual hypotheses that align to each other, compiled from the entire hypothesis list, where the weights are the hypothesis-level scores from the recognizer.

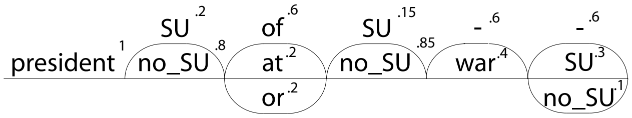

Figure 2: Confusion network for a merged hypothesis.

Finally, the best decision at each point is selected by choosing the words and boundaries with the highest probability. Here, the words and SUs are selected independently, so that we obtain the same words as would be selected without inserting the SU tokens and guarantee no degradation in WER. The key improvement is that the SU detection is now a result of detection across all recognizer hypotheses, which reduces the effect of word errors in the top hypothesis.

## 5 Experiments

Table 1 shows the results in terms of slot error rate on the four test sets. The middle column indicates the performance on a single hypothesis, with the words derived from the pruned set of N-best hypotheses. The right column indicates the performance of the system using multiple hypotheses merged with confusion networks.

Multiple hypotheses provide a reduction of error for both test sets of CTS (significant at $p<.02$ using the McNemar test), but give insignificant (and mixed) results for BN. The small increase in error for the BN evaluation set

|  | WER | SU error rate |  |
|---|---|---|---|
|  |  | Single Best | Confusion Nets |
| BN Dev | 12.2 | 55.79% | 54.45% |
| BN Eval | 12.0 | 57.78% | 58.42% |
| CTS Dev | 23.6 | 44.14% | 42.72% |
| CTS Eval | 22.2 | 44.95% | 44.01% |

Table 1: Word and SU error rates for single best vs. confusion nets.

may be due to the fact that the 1-best parameters were tuned on different news shows than were represented in the evaluation data.

We expected a greater gain from the use of confusion networks in CTS than BN, given the previously shown impact of WER on 1-best SU detection. Additionally, incorporating a larger number of N-best hypotheses has improved results in all experiments so far, so we would expect this trend to continue for additional increases, but time constraints limited our ability to run these larger experiments. One possible explanation for the relatively small performance gains is that we constrained the confusion network topology so that there was no change in the word recognition results. We imposed this constraint in our initial investigations to allow us to compare performance using the same words. It it possible that better performance could be obtained by using confusion network topologies that link words and metadata.

A more specific breakout of error improvement for the CTS development set is given in Table 2, showing that both recall and precision benefit from using the N-best framework. Including multiple hypotheses reduces the number of SU deletions (improves recall), but the primary gain is in reducing insertion errors (higher precision). The same effect holds for the CTS evaluation set.

|  | Single Best | Confusion Nets | Change |
|---|---|---|---|
| Deletions | 1623 | 1597 | -1.6% |
| Insertions | 872 | 818 | -6.2% |
| Total | 2495 | 2415 | -3.2% |

Table 2: Errors for CTS development set

## 6 Conclusion

Detecting sentence structure in automatic speech recognition provides important information for language processing or human understanding. Incorporating multiple hypotheses from word recognition output can improve overall detection of SUs in comparison to prediction on a single hypothesis. This is especially true for CTS, which suffers more from word errors and can therefore benefit from considering alternative hypotheses.

Future work will involve a tighter integration of SU detection and word recognition by including SU events directly in the recognition lattice. This will provide opportunities to investigate the interaction of automatic word recognition and structural metadata, hopefully resulting in reduced WER. We also plan to extend these methods to additional tasks such as disfluency detection.

### Acknowledgments

This work is supported in part by DARPA contract no. MDA972-02-C-0038, and made use of prosodic feature extraction and modeling tools developed under NSF-STIMULATE grant IRI-9619921. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these agencies.

## References

L. Breiman et al. 1984. *Classification And Regression Trees*. Wadsworth International Group, Belmont, CA.

H. Christensen, Y. Gotoh, and S. Renals. 2001. Punctuation annotation using statistical prosody models. In *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*.

J. Huang and G. Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *Proc. Eurospeech*.

J.-H. Kim and P. Woodland. 2001. The use of prosody in a combined system for punctuation generation and speech recognition. In *Proc. Eurospeech*, pages 2757–2760.

Y. Liu, E. Shriberg, and A. Stolcke. 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. In *Proc. Eurospeech*, volume 1, pages 957–960.

Y. Liu et al. 2003. MDE Research at ICSI+SRI+UW, NIST RT-03F Workshop. http://www.nist.gov/speech/tests/rt/rt2003/fall/presentations/.

L. Mangu, E. Brill, and A. Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, pages 373–400.

NIST. 2003. RT-03S Workshop Agenda and Presentations. http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/.

E. Shriberg et al. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, September.

A. Srivastava and F. Kubala. 2003. Sentence boundary detection in arabic speech. In *Proc. Eurospeech*, pages 949–952.

A. Stolcke. 2002. SRILM—an extensible language modeling toolkit. In *Proc. ICSLP*, volume 2, pages 901–904.

S. Strassel, 2003. *Simple Metadata Annotation Specification V5.0*. Linguistic Data Consortium.