

ACCOUNTING FOR STT UNCERTAINTY IN MDE

Dustin Hillard Mari Ostendorf

{hillard,mo}@ee.washington.edu
University of Washington
Department of Electrical Engineering
Seattle WA 98105

Andreas Stolcke

{stolcke}@speech.sri.com
SRI International
Speech Technology and Research Laboratory
Menlo Park CA 94025

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, HMM and maximum entropy models are 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.

1. INTRODUCTION

The output of most current automatic speech-to-text (STT) 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 describe 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 15-20%. These errors limit the effectiveness of sentence boundary prediction, because they introduce incorrect words to the word stream. Prior evaluations showed that 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 [1]. Including additional recognizer hypotheses allows for alternative word choices to inform sentence boundary prediction.

Our approach builds on work described in [2], but the method has been extended to handle multiple SU models with associated improvements to feature extraction for multiple STT hypotheses. To integrate the information from different alternatives, we first predict sentence boundaries in each hypothesized word sequence separately with two mod-

els: an HMM structure that integrates prosodic features in a decision tree with hidden event language modeling, and a maximum entropy (maxent) model that uses features related to those used in the HMM. 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 a baseline system and task domain reviewed in Section 2. Our approach integrates prediction on multiple recognizer hypotheses using confusion networks, as outlined in Section 3. Experimental results are detailed in Section 4, and the main conclusions of this work are summarized in Section 5.

2. 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 [3], which we refer to as the “V6” 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.

2.1. Baseline System

The automatic speech recognition system used was an updated version of that used by SRI in the Fall 2004 RT evaluations [4], with a WER of 18.6% on the CTS evaluation test set. The system performs multiple recognition and adaptation passes, and eventually produces 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 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 [5] from which the hypothesis with lowest expected word error is extracted. A confusion network is a compacted representation of a word lattice or N-best list, where the complexity of the lattice or list representation is reduced to a simpler form that maintains all possible paths (and more), transforming the space to a series of slots that each have word hypotheses (and null arcs) and associated posterior probabilities. 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 work on sentence boundary detection using lexical and prosodic features [6]. The system takes as input alignments from either reference or recognized (1-best) words, and combines lexical and prosodic information using multiple models. 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 using bagging technique to reduce the variability due to a single tree. Language models include word and class n-grams. The prosody and language model are combined using an HMM. In addition, a recent improvement is to combine prosody tree predictions with language cues using other modeling frameworks, such as maxent models and CRFs. For this work we include the maxent model, but not the CRF. A complete description of the most recent system advances is described in [7].

There are 40 hours of conversations available for training from the Switchboard corpus, 6 hours of development, and 3 hours evaluation test data drawn from both the Switchboard and Fisher corpora. Additional training data is available from 2003, but the annotations used a older guideline (V5), which differed significantly, so that data is not used in training for this year's CTS system. The evaluation set has roughly 5000 SUs.

2.2. Evaluation

The system is evaluated for conversational telephone speech (CTS) using training, development and test data annotated according to the V6 standard. The test data is that used in the DARPA Rich Transcription (RT) Fall 2004 evaluations.

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. (Substitution errors are included only when subtype is scored, and our focus is

on simple SU detection because subtype is determined in a subsequent detection stage, i.e. after the N-best SU detection.) When recognition output is used, the words generally do 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 md-eval (v18) scoring tool from NIST. For reference, the SU error rate for the baseline system using the 1-best output from the SRI recognizer is 40.63%.

3. 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.

3.1. 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 98% of the posterior mass (or to a maximum count of 1500), which gives comparable WER compared to the unpruned case.

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. For lexical features that span the boundaries of the recognizer segmentation, there is the potential to consider all of the N-best hypotheses. Since this quickly becomes very costly, we extend the current hypothesis to the left and right using the 1-best hypothesis from neighboring segments.

After obtaining the prosodic and lexical features, the HMM and maxent systems each predict 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.

3.2. Combining Hypotheses via Confusion Networks

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 arcs, as shown in Figure 1. (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 or maxent model. 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 STT system that generated that hypothesis.

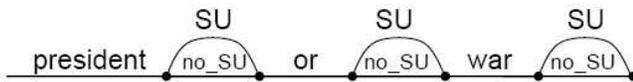


Fig. 1. Confusion network for a single hypothesis.

The confusion networks for each hypothesis are then merged with the SRI Language Modeling Toolkit [8] 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, as shown in Figure 2, contains slots with the word hypotheses from the N-best list and slots with the SU/no_SU arcs as before but with new confidence estimates. 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 scaled hypothesis-level posteriors.

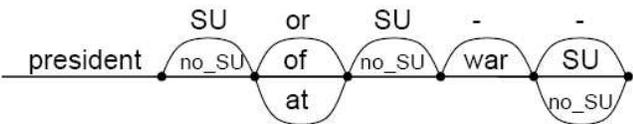


Fig. 2. Confusion network for a merged hypothesis.

A combined confusion network is produced for each model type: HMM and maxent. The same alignment and combination is then applied to obtain a final confusion network. The HMM and maxent confusion networks are each given equal weight for this final combination.

Finally, the best decision at each point is selected by choosing the words and boundary events 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.

4. EXPERIMENTS

Table 1 shows the results in terms of slot error rate on the RT04 CTS Evaluation Set. 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. There is an improvement of 0.7% absolute in the SU detection score, which is significant at $p < .08$ using a matched pair test on segments defined by large pauses [9]. This improvement also translates directly to an improvement in the subtype classification score, i.e. when substitution errors are included the error rate is 52.2% for the pruned 1-best system and 51.5% for the confusion network system. However, examining the performance of the two systems over a range of operating points via a decision-error tradeoff (DET) curve (Figure 3) shows that there is not a consistent advantage to the N-best system.

	SU error rate	
	Single Best	Conf. Nets
CTS HMM	42.0%	41.9%
CTS Maxent	42.4%	43.2%
CTS HMM+Maxent	41.2%	40.5%

Table 1. SU error rates for single best vs. confusion nets using N-best list pruning.

The negative results obtained here are in contrast to earlier work [2], which showed somewhat greater benefit to the use of confusion networks. There have been some changes to the task definition (V5 vs. V6 specification) and there are always statistical variations across different test sets, but more importantly the SU detection system has improved. One possible reason for the smaller benefit in the current work is that the N-best framework cannot take advantage of a new turn labeling strategy used in the 1-best system. We also note that the introduction of DET-based error reporting [9] allows us to see that there may be greater or lesser advantages at different operating points.

Though gains were bigger in prior work, they were still relatively small. At that time, we hypothesized that N-best pruning was limiting performance, but here we find that the

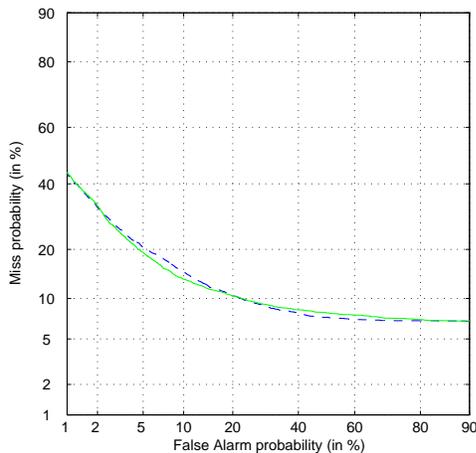


Fig. 3. DET curve illustrating SU detection error trade-offs for pruned 1-best (solid line) vs. confusion network (dashed line) decoding.

WER is similar with and without pruning so that cannot explain the negative results. Since improvement in WER clearly has a significant impact on SU detection, and since the N-best oracle error rate is significantly better than the 1-best WER (roughly 30-40% lower), we hypothesize that either the MDE confidence needs to be more tightly linked to the word hypothesis and/or that improved word confidence estimates are needed. The current paradigm treats the SU events in the confusion network as conditionally independent of the word context in an attempt to retain the same 1-best word sequence after SU processing, though the posterior weights are in fact derived from the language cues. It may be that explicitly representing the SU coupling will help the SUs without hurting the WER.

5. CONCLUSIONS

Detecting sentence structure in automatic speech recognition provides important information for automatic language processing and human understanding. Incorporating multiple hypotheses from word recognition output can improve overall detection of SUs in comparison to prediction on a single hypothesis, although only small gains have been realized so far. Different MDE modeling approaches influence the benefits of using multiple hypotheses, but we hypothesize that decoupling the SU and word events is probably the biggest limitation of the current framework.

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

The authors are grateful to Yang Liu for providing the SU detection system, and also to Liz Shriberg for consulting on long-distance prosodic feature extraction approximations for N-best processing. 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.

6. REFERENCES

- [1] National Institute of Standards and Technology, "RT-03F workshop agenda and presentations," <http://www.nist.gov/speech/tests/rt/rt2003/fall/presentations/>, 2003.
- [2] D. Hillard, M. Ostendorf, A. Stolcke, Y. Liu, and E. Shriberg, "Improving automatic sentence boundary detection with n-best recognition results and confusion networks," in *Proc. HLT-NAACL*, May 2004, pp. 69–72.
- [3] S. Strassel, *Simple Metadata Annotation Specification V6.2*, Linguistic Data Consortium, 2004.
- [4] National Institute of Standards and Technology, "RT-03S workshop agenda and presentations," <http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/>, 2003.
- [5] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, pp. 373–400, 2000.
- [6] Y. Liu *et al.*, "The ICSI-SRI-UW metadata extraction system," in *Proc. ICSLP*, October 2004, vol. I, pp. 577–580.
- [7] Y. Liu *et al.*, "The ICSI-SRI-UW metadata extraction system," in *Proc. of the NIST RT04 Workshop*, November 2004.
- [8] A. Stolcke, "SRILM—an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, vol. 2, pp. 901–904.
- [9] D. Hillard and M. Ostendorf, "Scoring structural MDE: Towards more meaningful error rates," in *Proc. of the NIST RT04 Workshop*, November 2004.