Attention Shift Decoding for Conversational Speech Recognition

Raghunandan Kumaran, Jeff Bilmes, Katrin Kirchhoff

Department of Electrical Engineering, University of Washington, Seattle, WA 98195
{rkumaran,bilmes,katrin}@ee.washington.edu

Abstract

We introduce a novel approach to decoding in speech recognition (termed attention-shift decoding) that attempts to mimic aspects of human speech recognition responsible for robustness in processing conversational speech. Our approach is a radical departure from traditional decoding algorithms for speech recognition. We propose a method to first identify reliable regions of the speech signal and then use these to help decode the unreliable regions, thus conditioning on potentially non-consecutive portions of the signal. We test this approach in a second-pass rescoring framework and compare it to standard second-pass rescoring. On a conversational telephone speech recognition task (EARS RT-03 CTS evaluation), our approach shows an improvement of 2.6% absolute when using oracle information for detecting the reliable regions, and 0.4% absolute when detecting the reliable regions automatically.

Index Terms: speech decoding, speech recognition

1. Introduction

Conversational speech is notorious for extreme pronunciation variability, caused by e.g. articulatory reduction and coarticulation. Not surprisingly, context plays a key role for processing conversational speech. Classic studies of speech intelligibility [1] have shown that words excised from running speech are often unintelligible unless they are heard in context. In our own informal listening experiments on conversational telephone speech corpora (Switchboard and Fisher), we found that even isolated segments as long as five consecutive words were sometimes impossible for human listeners to identify unless the entire utterance was available as context. The notion that humans use wide-ranging cues during speech processing, including those occurring later than the current word, is corroborated by clinical studies. E.g. [2] studied aphasia or stroke patients that had difficulties in distinguishing similar-sounding words in everyday conversational speech (much like speech recognizers!) but showed excellent performance in isolated-word discrimination. It was found that their sentence processing difficulties resulted from an impairment in phonological short-term memory, i.e., a stored buffer (1-2 seconds long) of partial phonological analysis. It seems that healthy listeners use this buffer to re-evaluate earlier input during on-line sentence-processing in order to resolve lexical ambiguities. In other words, humans seem to focus on prominent, acoustically salient regions of the signal first and then fill in or re-evaluate the “holes” using contextual information in combination with stored partial phonological information.

The current decoding paradigm in speech recognition does not make use of such strategies. In a standard HMM-based system, speech is decoded by finding the assignment $h^*_{1:T}$ that maximizes $p(h_{1:T}, \bar{o}_{1:T})$, where $h_{1:T}$ is a sequence of T state values, and $\bar{o}_{1:T}$ is a sequence of T observation vectors. The HMM factorization properties allow us to perform this maximization efficiently using dynamic programming, which typically consists of running a left-to-right pass over the data followed by a right-to-left pass (we refer specifically to the popular synchronous decoding paradigm). Since the state space is enormous for large-vocabulary ASR system, pruning is necessary during (at least) the first pass, but such pruning schemes are ignorant of the current reliability of the speech signal – i.e., this approach is likely to prune away too few hypotheses during a reliable region, when much more aggressive pruning could be utilized. Similarly, it might prune away too many hypotheses during unreliable regions at which time it is important to maintain a wide diversity of hypotheses. It is true that systems typically make use of multiple recognition passes. Even so, significant benefits still could be achieved if a system, when it is making such pruning decisions, was informed about regions nearby in time where the speech was known to be reliably (if not perfectly) recognized.

Drawing on these insights into human and machine speech processing, we propose a different strategy: initial word hypotheses should be produced on the basis of the most reliable parts or *islands* of the speech signals. Further recognition processing will then work *inwards* from these non-consecutive regions and fill in the intermediate parts or *gaps* using more specialized models and a variety of information sources. Similar “island-based” approaches have been proposed in other fields (e.g. parsing or handwriting recognition) and were pursued in a rule-based fashion in the early days of ASR but have never been integrated into a state-of-the-art statistical framework. We are pursuing a long-term project to develop a viable island-based statistical decoding framework for ASR; in this paper, we present a pilot study to test the promise of this approach. To this end, in this paper we implement island-based rescoring whereby first-pass lattices obtained from a standard large-vocabulary ASR system are rescored using island/gap information.

2. Attention Shift Decoding

The attention-shift decoding approach focuses first on the most reliable portions of the speech signal. While the “reliable” portion of speech might be relative to a given speech recognizer, our approach acknowledges this fact by defining the reliable portion of speech with respect to a given recognition system. While this might seem similar to the general idea of boosting [3], our approach is quite different. In attention-shift decoding, the hypotheses obtained for these regions are used in combination with evidence from the remaining, less reliable, regions to generate word hypotheses for the latter (see Figure 1). Thus,

This work was supported by an ONR MURI grant, No. N000140510388

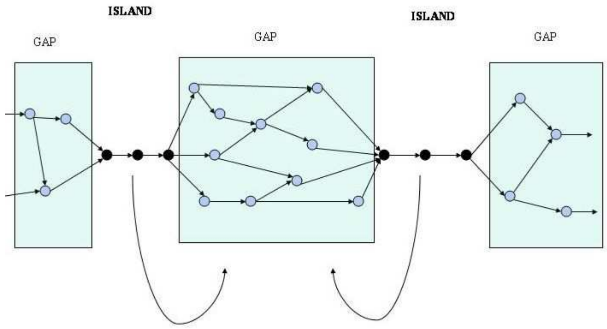

Figure 1: *Attention-Shift Decoding*

there is no initial either left-to-right or right-to-left pass during which uninformed pruning must be performed. Instead, the final sentence hypothesis iteratively evolves by combining evidence from potentially nonconsecutive portions of the speech signal. The Hearsay speech recognition system [4] was an early speech recognition system based on a similar idea; however, it was rule-based and did not make use of state-of-the-art statistical and machine learning frameworks. The basic idea of using reliable regions of a signal to determine unreliable portions of a signal has also been used in other areas, such as parsing [5] and handwriting recognition [6]. The present work is part of a long-term project designed to explore the concept of non-consecutive "island-based" decoding in combination with a rigorous statistical framework.

A first-pass decoding framework for this approach is currently under development. In this paper, we test its basic premises in second-pass rescoring. In standard rescoring, scores in the original lattice are replaced with new scores (or new scores are added); the best path is then determined from the weighted combination of all scores. Instead of a lattice, an N-best list or confusion network representation can be used. In our framework, we use information about the correct vs. incorrect word hypotheses in the 1-best output to constrain the search space, pruning away the hypotheses competing with correct words and leaving the competing hypotheses for incorrect words intact. Only the latter are thus influenced by the new scores. To easily identify competing hypotheses, the lattices are first converted into a confusion network [7] - this collapses similar hypotheses (e.g. those with identical word labels and slightly different time stamps) into a single one and displays alternative hypotheses for the same time interval. In determining correct vs. incorrect regions (islands vs. gaps) we run two different experiments: one that uses oracle information, in order to determine the maximum potential improvement from our method, and one that uses the output of an automatic gap vs. island classifier. At this point, hard decisions (correct/incorrect) are used rather than the classification score; at a later point, we will investigate the use of soft version of this.

3. Island vs. Gap Classification

The problem of island vs. gap classification is similar (though not identical) to the problem of word confidence estimation in speech. The higher the confidence of a word in the lattice, the more likely it is part of an island segment and vice-versa. However, one important difference between word confidence estimation and our task is that we are more concerned about the accuracy with which gap regions are identified. The reason is that our method relies quite heavily on correctly-identified island regions for rescoring the gap regions – thus, if a wrong word is used as an island, then the gap region will be rescored based on wrong or unreliable information. Our classifiers are therefore tuned to achieve high accuracy on gaps.

Word confidence estimation is a well-investigated problem in speech recognition. Typical approaches use a collection of word features, including features of adjacent words, and then apply a statistical classifier to map each feature vector to a word confidence score. Features based on the acoustic model, the language model, the decoding process and word semantics have been proposed. Classifiers investigated include linear discriminant analysis followed by linear thresholds, Bayes classifiers, neural networks, SVMs, boosted classifiers etc. [8] summarizes most of the research related to word confidence estimation.

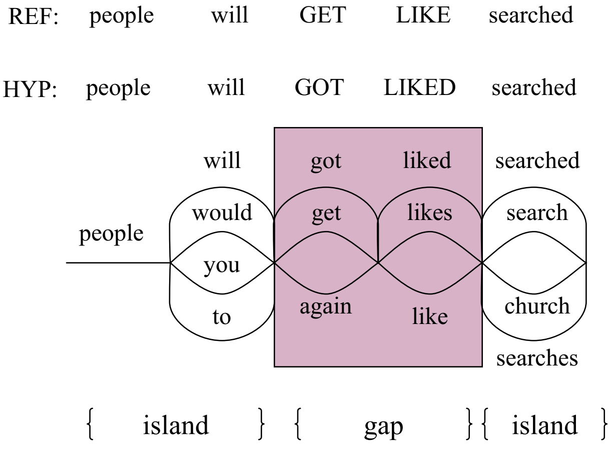
REF: people will GET LIKE searched
HYP: people will GOT LIKED searched
people
will
would
you
to
got
get
again
liked
likes
like
searched
search
church
searches
{ island } { gap }{ island }

Figure 2: *Confusion network showing the island and gap regions*

3.1. Features

The feature set we use has been derived by augmenting some of the more popular feature sets for confidence estimation with phone-based, acoustic-prosodic and linguistic features. Fig. 2 shows a typical confusion network. The words in each "slot" (ensemble of arcs between the same two nodes) of the confusion network are ranked according to the normalized posterior score, and the top-ranking words from each slot form the 1-best sentence hypothesis. The figure also shows the island and gap regions that have been identified by aligning the 1-best hypothesis from the confusion network with the reference hypothesis, which is an example of the oracle experiment described later. The following basic features are obtained for the top word in the each slot of the confusion network: normalized posterior probability, acoustic score, language model (LM) score, word length in number of phones, word duration in seconds, the number of competing word hypotheses in the slot, the difference in posterior probability between the first-best and second-best hypotheses in the slot, and the entropy of the probability distribution over all words in the slot.

We also used phone-based features, viz. distance scores (penalizing dissimilar phones) for neighbouring phones and edit distance for phone strings of competing hypotheses. The former is obtained from phonetic considerations of similarity of place and manner of articulation; the latter is the standard Levenshtein distance. In particular, we use *phone distance left*: The phone distance between the last phone of the top word in the slot to the left and the first phone of the top word in the current slot. *phone distance right*: The phone distance between the last phone of the top word in the current slot and the first phone of the top word in the slot to the right. *top-two phone edit distance*: edit distance in terms of phones between the top word in the slot and the second word in the slot.

In addition, we used acoustic-prosodic features. Features based on energy and F0 have been used in many applications like pitch accent detection, prominence detection etc. (e.g. [9]). Research on the perception and detection of prosody has shown that acoustic features based on F0, duration, and intensity are all indicators of prosodic prominence. Prominent words are often likely to be recognized correctly, making these features potentially helpful for island-gap classification. Each utterance is mean-variance normalized and then the acoustic signal between the start and end times of each word is used for extracting the features. The start and end times for the words are obtained from the information in the confusion network. The features extracted are: *normalized energy, maximum F0, minimum F0, average F0, slope of the F0 curve*, and *F0 excursion*, i.e. log of the ratio of maximum F0 over minimum F0.

The final group of features is linguistically motivated and includes: *function/content word*: A binary feature (1 if the word is a function word and 0 if it is a content word). The motivation for this feature is that content words are more likely to be recognized correctly than function words (unless it is an out of vocabulary word). *POS tags*: A standard POS tagger (maxEnt tagger [10]) trained on the Switchboard data set was used to tag the top hypothesis in the confusion network. An indicator feature vector which is 1 for the POS tag of the word, while the other possible POS tags are 0, is used to represent this feature.

3.2. Classifiers

We tried three of the most common classifiers used in word confidence estimation namely boosted decision trees, SVMs and neural nets. Boosting has become a widespread technique for improving weak learners (see [3]). The weak learner we have used in here is a simple decision tree with three nodes. The decision tree uses one feature for classification at each node. We experimented with different tree configurations and found that a tree with three nodes performed the best when compared to a decision stump or higher number of nodes. The SVM classifier we used is a linear SVM using SVMLin [11], which is well-suited for classification problems involving a large number of examples and features.(We could not use other kernels because of the size of the data set). The neural net classifier was implemented using Quicknet. The input layer had 78 units, hidden layer 30 units and 2 output units. Softmax was used as the activation function.

4. Data and Experiments

The experiments reported here were performed on conversational telephone speech (Switchboard and Fisher data). The development and test sets are those used for the EARS RT-03 CTS evaluation. We use lattices provided by SRI. Lattices were generated by one of the 3 final decoding passes in the SRI CTS system, using cross-word triphone PLP acoustic models and a multiword 3-gram SuperARV language model. The resulting lattices were then rescored with a 4-gram SuperARV LM and pronunciation scores, which are all recorded in the lattices. The dev set lattices are used for training our island-gap classifiers and optimizing language model scaling factors; the eval set lattices are used for testing. There are 36 speakers and 72 conversations each in the dev set and eval sets, with 2930 segments in the dev set and 2910 segments in the eval set.

For obtaining the confusion networks of these lattices, as well as rescoring them, SRI's lattice-tool was used. The language model used for rescoring these confusion networks is a 4-gram backoff LM, trained using about 20M words from Fisher data set, 3M words from Switchboard and about 102M words from Fisher-related web-data [12] and modified Kneser-Ney smoothing. The vocabulary was chosen as the 64,000 most frequent words in the trainig data. The LM had a perplexity of 65.27 on the dev set references. Table 1 shows the result of standard LM rescoring of the confusion networks, i.e. the original language model scores were replaced with the new LM scores, which were then used in combination with the posterior scores to choose the 1-best hypothesis. The LM weight was reoptimized on the dev set. Island-Gap information was not used. Note that there is no improvement in the word error rate (WER) on the dev set, and only 0.1% improvement on the eval set. This may be because the original lattice contains LM scores from a language model that is already highly optimized and includes not just n-gram but also grammar information.

WER	Dev set	Eval set
Baseline	27.1	25.1
Rescoring	27.1	25.0

Table 1: *Result of standard rescoring of confusion networks using the 4-gram LM.*

4.1. Oracle rescoring experiments

We first conducted oracle experiments, using perfect information about islands and gaps. The 1-best hypothesis from the confusion networks are aligned with the reference transcripts; those words that do not match the reference are marked as 'gaps' while correct words are marked as 'islands'. For the words in the 1-best hypothesis that belong to the island region, only the top word in the corresponding slot of the confusion network is retained while the rest of the alternate hypotheses for that slot are eliminated. Each island-gap-island segment is treated as a separate confusion network and is rescored using the LM. The 1-best from this confusion network is obtained by using a combination of the new LM score and the normalized posterior score, both weighted with scaling factors optimized (using grid search) on the dev set.

WER	Dev set	Eval set
Baseline	27.1	25.1
Attention-shift decoding	24.3	22.5

Table 2: *Oracle results for attention-shift decoding.*

Table 2 shows the oracle results. We see that even though the LM did not give any improvement in standard rescoring, there is a significant improvement of 2.8% on the dev set and 2.6% on the eval set by using the oracle information. This shows that given very good knowledge about the gap and island regions in the confusion network, we can achieve significant improvements in recognition performance. To study the extent of the influence the island region has on the gap region, we experimented with different gap lengths, in each case fixing the island portion to be 1 word long on either side. Fig 3 shows the improvement in WER for different gap lengths. We see that

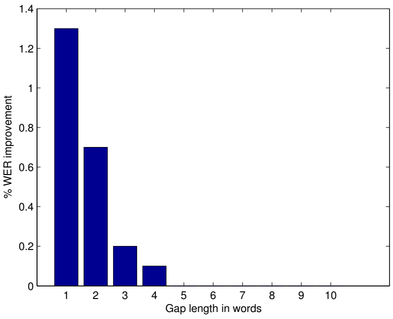

Figure 3: *WER improvement vs gap length.*

the improvement is largest for a gap of 1 word and falls off exponentially as the gap length increases. It was also observed that most of the words that were corrected were function words. Beyond a gap length of 4, there is no improvement, probably due to the limit of the n-gram model to order 4. Higher-order LMs were tried but did not show any additional improvement.

4.2. Island-Gap classification experiments

Classifier	Class 0 accuracy	Class 1 accuracy	Overall accuracy
Boosting	59.21%	82.54%	62.59%
SVM	55.09%	82.33%	59.64%
neural-net	60.85%	80.58%	63.47%

Table 3: *Classification results. Class 0 = correct words, Class 1 = incorrect words.*

For training the island-gap classifiers, the confusion networks of the dev set were used, while for testing those of the eval set were used. The 1-best hypotheses from the confusion networks of the dev set were aligned with the references to identify correct and incorrect words (which correspond to the two classes on which the binary classifier is trained). The data set was balanced such that both classes were represented with approximately equal priors. The features with a very large dynamic range such as the LM score, acoustic score, F0 features, etc. were normalized across all samples; moreover the feature vector for each sample was normalized. Table 3 shows the classification results of the three classifiers. All three classifiers have roughly the same performance, with the boosted decision tree having the best accuracy on the gap words (incorrect words). Some of the features that were chosen during the boosting iterations were: posterior score, posterior score difference, number of alternative word hypotheses per slot, LM score, segment entropy, number of phones, F0 excursion and energy.

4.3. Non-oracle experiments

WER	Dev set	Eval set
Baseline	27.1	25.1
Attention-shift decoding	26.7	24.7

Table 4: *Non-oracle rescoring results.*

For the non-oracle experiments we use automatically detected island-gap regions. The dev set data is used to train classifiers as described in section 3 to identify island and gap regions. Rescoring then proceeds in the same manner as in the oracle case. Table 4 shows the results from the non-oracle experiments. The improvement in WER on both the dev and eval sets is 0.4% absolute which is considerable given the fact these lattices have already been through 2 rescoring passes using more complex knowledge sources than just a 4-gram LM. However, the improvement is not as much as we see in the oracle case, which is mainly due to the fact that the classifier is not as accurate in identifying the island-gap regions as desired.

5. Conclusions

We have described an algorithm that makes use of reliability/unreliability labels for second-pass rescoring and have shown that our method performs significantly better than standard rescoring methods if oracle information is used. Beyond this proof-of-concept experiment, we have used automatically obtained labels and demonstrate a small improvement, although not as large as that obtained in the oracle case. Our future work will concentrate on improving island vs. gap detection (e.g. by using parse-based features) and on integrating this approach into first-pass decoding.

Acknowledgments
We gratefully acknowledge SRI for providing the lattices.

6. References

[1] I. Pollack and J. Pickett, "The intelligibility of excerpts from conversation," *Language and Speech*, vol. 6, pp. 165–171, 1963.

[2] M.C. Silveri and A. Cappa, "Segregation of the neural correlates of language and phonological short-term memory," *Cortex*, vol. 39, pp. 913–925, 2003.

[3] Robert E. Schapire, "A brief introduction to boosting," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.

[4] L.D. Erman, F. Hayes-Roth, V.R. Lesser, and D. R. Reddy, "The Hearsay-II speech understanding system: Integrating knowledge to resolve uncertainty," *ACM Comput. Surv.*, vol. 12(2), pp. 213–253, 1980.

[5] A. Corazza, R. De Mori, R. Gretter, and G. Satta, "Stochastic context-free grammars for island-driven probabilistic parsing," in *Proceedings of International Workshop on Parsing Technologies*, 1991.

[6] J. Pitrelli, J. Subrahmonia, and B. Maison, "Toward island-of-reliability-driven very-large-vocabulary on-line handwriting recognition using character confidence scoring," in *Proc. ICASSP*, 2001.

[7] 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*, vol. 14(4), pp. 373–400, 2000.

[8] Hui Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, 2005.

[9] J. M. Brenier, D. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Proceedings of EUROSPEECH*, 2005.

[10] Adwait Ratnaparkhi, "A linear observed time statistical parser based on maximum entropy models," in *Second Conference on Empirical Methods in Natural Language Processing*, 1997.

[11] V. Sindhwani and S. S. Keerthi, "Large scale semi-supervised linear SVMs," in *Proceedings of SIGIR*, 2006.

[12] Ivan Bulyko, Mari Ostendorf, and Andreas Stolcke, "Class-dependent interpolation for estimating language models from multiple text sources," Tech. Rep. UWEETR-2003-0003, University of Washington, 2003.