Continuous Speech Recognition Using Attention Shift Decoding with Soft Decision

Ozlem Kalinli and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, CA, USA
kalinli@usc.edu, shri@sipi.usc.edu

Abstract

We present an attention shift decoding (ASD) method inspired by human speech recognition. In contrast to the traditional automatic speech recognition (ASR) systems, ASD decodes speech inconsecutively using reliability criteria; the *gaps* (unreliable speech regions) are decoded with the evidence of *islands* (reliable speech regions). On the BU Radio News Corpus, ASD provides significant improvement (2.9% absolute) over the baseline ASR results when it is used with oracle island-gap information. At the core of the ASD method is the automatic island-gap detection. Here, we propose a new feature set for automatic island-gap detection which achieves 83.7% accuracy. To cope with the imperfect nature of the island-gap classification, we also propose a new ASD algorithm using soft decision. The ASD with soft decision provides 0.4% absolute (2.2% relative) improvement over the baseline ASR results when it is used with automatically detected islands and gaps.

Index Terms: speech recognition, decoding, attention, island.

1. Introduction

Human-like speech processing has been an inspiration and motivation for researchers for many years to improve the performance of computational models and machine processing applications. Humans can successfully recognize speech with high accuracy despite conditions such as highly variable speaking styles, noise conditions, overlapping sources, etc. In contrast, the machine performance typically degrades drastically in such conditions. Existing automatic speech recognition (ASR) systems have modeled some parts of the human speech recognition process and found them to be beneficial; signal processing in the peripheral auditory system is a good example. There are however other possibilities that offer promise. One of those that can be considered within ASR systems is the "*attention*" mechanism human use.

Humans can precisely process and interpret complex scenes in real time despite the tremendous number of stimuli impinging the senses. One of the key enablers of this capability is the attention mechanism that selects a subset of available sensory information before fully processing all stimuli at once [1]. Only the selectively attended incoming stimuli are allowed to progress through the cortical hierarchy for high-level processing to recognize the details of the stimuli. Thus, it is believed that humans process a scene nonconsecutively in a selective way. In addition to this, the experiments in [2] have shown that words segmented from running speech are often unintelligible even for humans, and they become intelligible when they are heard in the context of an utterance. Also, the experiments in [3] showed that humans use a short-term memory buffer (about 1-2 sec long) which when injured causes sentence processing difficulty. These experiments indicate that i) humans use context information while decoding speech, ii) humans use a buffer that stores a string of words while recognizing individual words within a sentence. Based on the attention theory and supporting experimental findings, it is believed that humans first process and recognize salient or prominent parts of speech. Then they finalize recognition of non-salient parts of speech using the contextual information together with their segmental properties.

Prior research that has focused on the notion of attention in speech understanding dates back to Hearsay system [4]. The Hearsay speech understanding system is one of the early works which proposed to resolve uncertainty using many knowledge sources in a selective structure [4]. One of the limitations of the Hearsay system is that it was rule-based and it has not been implemented within the state-of-the art machine learning framework. Human like non-consecutive speech recognition has been the motivation of some other work in the past. In [5], an island-driven continuous speech recognition system that uses word spotting and word verification was proposed. The system described in [5] first detects a noun as an island in a small vocabulary continuous speech and then expands the island by verifying neighboring words predicted by a word pair grammar until all parts of speech were filled. Island-driven search technique has been applied to handwriting recognition in [6, 7], and parsing in [8]. However, [6, 7, 8] followed a different approach than [5] and used reliable parts of signal (called as *island*) to determine unreliable parts of signal (called *gaps*). Recently, the idea of island of reliability driven search was applied to continuous speech recognition in [9], and it was concluded that the speech recognition performance was highly dependent on the accuracy of automatic detection of islands of continuous speech.

In this work, we explore the possibility of improving automatic speech recognition performance by using a human like *attention shift decoding* (ASD) approach. The presented method builds on the ideas proposed in [6, 9]. The method first finds the islands of continuous speech, and then recognizes them. The islands consist of reliable regions of speech for an automatic speech recognizer. Then, the islands are expanded by verifying the neighboring words using a statistical language model within a lattice search algorithm. Thus, the algorithm uses neither left-to-right nor right-to-left consecutive search paradigm as in the conventional ASR systems. It starts decoding from the islands of the speech and then fills in the gaps using the contextual information to make a selection amongst the word hypotheses obtained from the segmental features.

The main contributions of the paper are as follows: as mentioned earlier, the performance of an attention shift decoding algorithm highly depends on the automatic detection of islands in continuous speech with high accuracy. Hence, one of the main focuses of the paper is to explore the parameters that will lead us to achieve high island detection accuracy. Here, we propose a new set of features that is inspired by both human and machine recognition of speech for detection of islands. In addition to this, we propose a novel attention shift decoding method using soft decision to cope with the imperfect nature of island detection. Finally, we present continuous speech recognition experiments and results with attention shift decoding using both soft and hard decision for completeness.

This research was supported by DARPA, ONR and Army.

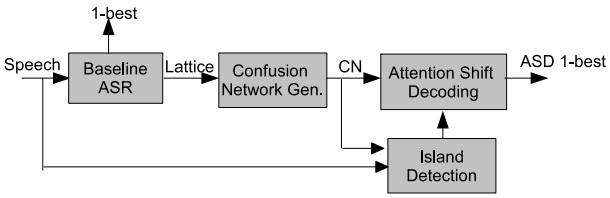

Figure 1: Block diagram of Attention Shift Decoding Method

The paper is organized as follows: the ASD method is explained in Section 2 followed by the automatic island detection in Section 3. The experimental results and conclusions are presented in Sections 4 and 5, respectively.

2. Attention Shift Decoding

Here, we present an attention shift decoding method that decodes speech nonconsecutively based on reliability criteria inspired by human speech recognition. The block diagram of the ASD system is shown in Fig. 1. The method first decodes each speech utterance using an automatic speech recognizer and provides a word lattice output in addition to the 1-best sentence hypothesis for each utterance. A word lattice may contain a large number of competing word hypotheses; hence they are transformed to word confusion networks (CN) to easily obtain the competing words for each time interval [10]. A word confusion network for a sample utterance together with its transcription (TRA) and the ASR 1-best output (HYP) is illustrated in Fig. 2. In a CN, the words in each time interval or slot (all of the arcs between two neighbor nodes) are sorted based on the normalized posterior probability as shown in Fig. 2. The top words from each time interval form the 1-best output of the ASR. Then, the correctly recognized words form *islands*, and the incorrectly recognized words form *gaps*. After identifying islands and gaps for an utterance, the gaps are filled by decoding the utterance with the evidence of neighboring islands. As mentioned before, we propose a novel ASD method using soft decision. For sake of clarity, we first present ASD method using hard decision [9].

2.1. ASD Using Hard Decision

The method first detects islands for an automatic speech recognizer, and finalizes the recognition of these reliable words by pruning out the alternative hypotheses for the island words. For example, in Fig. 2, in the second time interval only the top hypothesis (the arc that carries word *give*) will be kept by pruning out the other three word hypotheses since this is an island. In other words, the recognition of island words is finalized, and they cannot be altered in the later steps. At this stage, the hypotheses for the gaps are left intact. Next, the new pruned confusion networks are re-scored with a language model (LM). After re-scoring, the gap words carry new LM scores that are based on the island words; hence it is believed to be more accurate. Finally, the 1-best recognition output is obtained using the new LM scores together with the normalized posterior probabilities from the original confusion network since they are intact.

2.2. ASD Using Soft Decision

At the heart of the ASD method is the automatic island-gap detection, which is an inherently challenging problem. Usually the island-gap detectors are prone to errors, hence using a hard decision by taking the island-gap detector output as binary decision and pruning the confusion network accordingly may not benefit enough from ASD. Thus, we propose an alternative ASD scheme using soft decision to deal with the imperfect nature of automatic island detection.

The island-gap detector is designed such that it returns the posterior probability of the top word hypothesis in a time slot being island given the features; i.e., $P(I|F)$ where I is the island label, F is the features explained in Section 3. Then, for the other alternative word hypotheses in the same time slot, the probability of being island is computed as $1 - P(I|F)$; i.e., the more likely the top word is an island, the less likely the alternative words in the same slot can be an island (the correct word). We enrich the confusion networks by embedding a new score of island by modifying the standard ASR equation such as:

$$\mathbf{W}^* \approx \arg\max_{\mathbf{W}} P(\mathbf{A}|\mathbf{W})^{AS}.P(\mathbf{W})^{LS}.P(\mathbf{I}|\mathbf{F})^{IS} \quad (1)$$

where $\mathbf{W}$ stands for the word sequence, $P(\mathbf{A}|\mathbf{W})$ is the acoustic model score with scale AS, $P(\mathbf{W})$ is the language model score with scale LS, and IS is the island scale.

As discussed before this is a second-pass decoding, hence instead of acoustic model score the normalized posterior scales are used. Enriching the confusion network by adding an island score does nothing but re-ranking the hypotheses in each time slot based on the combined posterior and island scores; i.e., for simplicity consider that the posterior and the island scales are equal: $AS = IS$ in Eq. 1. In other words, if a first-best word has high probability of being island then its posterior score will be boosted, while the alterative words in the same slot will be penalized, otherwise the top word will be penalized while candidacy of the alternative words in the same slot will be promoted.

In our oracle experiments, where it is assumed that the islands and gaps are known perfectly, i.e., $P(I|F) = 1$ for the islands and $P(I|F) = 0$ for the gaps, the soft decision becomes similar to hard decision with one difference: for the islands the alternative words in the same time slot are going to be pruned as in the hard decision; for the gaps the top word hypothesis is going to be also pruned since $P(I|F) = 0$, while the remaining alternative words in a time slot will be left intact, which is different than the one in the hard-decision.

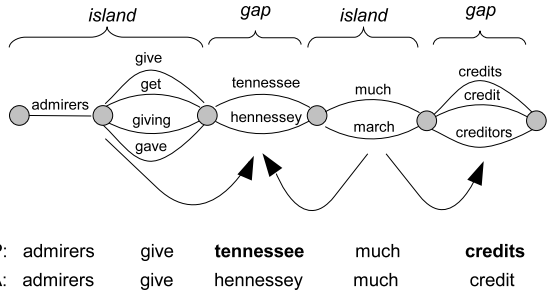

Figure 2: Sample word confusion network with islands and gaps

3. Automatic Island Detection

At the heart of the ASD method is the island detection. The goal is to detect whether the top word hypothesis in each time slot in a confusion network is island or not. Here, we propose a new set of features for the automatic island-gap detection inspired by both human and machine speech recognition. First, we summarize some key factors taking place in human word recognition which also lead us to select some features in our island-gap classifier. The references and a review of the research on spoken word recognition can be found in [11]. It is not surprising that there is some commonality between the factors affecting both human and machine recognition of speech. When it is applicable, these similarities are addressed as this section evolves.

Successful human communication depends on word recognition [11]. There is no doubt that segmental features provide information about which sounds are in an utterance. For fluent speakers of a language, the words are usually stored in long-term memory, and hence lexical access is an essential part of word recognition [11]. Segmental and suprasegmental information are extracted from the signal and used in lexical access to

activate a set of candidate words in lexicon [11].The factors affecting lexical access and word activation are as follows: It was found that segmental mismatch is more disruptive of lexical access in word initial than in word final position since words with initial mispronunciation have to recover from a poor start, while the words with final mispronunciation can be already highly activated before the mismatch occurs [11]. This is also valid for machine recognition of speech from the perspective of search paradigm. Hence, we compared the first (*f_syl*) and last syllable (*l_syl*) of the first-best and second-best word hypothesis in a time slot; i.e., if the first syllables of the top two words are the same then *f_syl*=1, otherwise *f_syl*=0.

Second, the mismatched segments in short words appear to be more disruptive than the ones in long words [11]. Hence, we used three word length measures to capture the information for the top word hypothesis in a time slot: word duration in milliseconds (*duration*) (obtained from the CN), the number of phones (*n_ph*) and syllables (*n_syl*) in the word. Syllabification software from NIST [12] is used for syllabifying words using their phoneme strings. The approximate phone duration (*ph_dur*) is also used (word duration in milliseconds divided by the number of phones). Third, lexical neighbors play a role in word recognition; the presence or absence of similar sounding words influences the effect of segmental mismatch [11]. Thus, we captured the distance information among the set of word hypotheses in a time slot. The phone based distance scores are computed using the standard Levenshtein distance. We computed minimum (*min_dist*) and maximum (*max_dist*) phone distance in a time slot after computing phone distances between all possible pairs of words in a time slot. We also computed the phone distance between the first-best and second-best hypothesis words in a time slot (compete distance: *comp_dist*), and the normalized compete distance, (*n_dist*: compete distance divided by the number of phonemes in the first-best word).

When the number of similar sounding words increases the word recognition becomes harder and gets delayed for humans [11]. Similarly, the number of word hypotheses within a time slot is also a clear indication of the level of difficulty an ASR is having. For example, it was observed that usually when there is an out-of-vocabulary (OOV) word in a sentence, the ASR system makes mistakes and usually the number of hypotheses in these time slots is larger. Hence, we used the number of alternative word hypotheses within a time slot (*ncomp*) to capture the number of candidate words for ASR.

The word frequency also affects lexical activation; i.e., humans recognize more accurately the words they use frequently [11]. This is also true for the ASR systems; the words which occur more frequently usually have more data samples during training hence may be recognized more accurately. From the LM, we used unigram (*unigram*) probability of the top hypothesis word in a time slot to capture the word occurrence frequency.

Suprasegmental information is another cue used during lexical access by humans. English listeners appear to be sensitive to whether a syllable is prominent (stressed) or not since they believe that content words in English tend to begin with prominent syllables [11]. Similarly, content words are more likely to be recognized correctly than function words in ASR. Hence, we used prominence (*prom*) of the top word hypothesis in a time slot as a feature for island-gap classifier. For this, we used our previously proposed top-down attention model that can detect the prominent words in speech with high accuracy from the acoustic signal [13]. As shown in Fig 1, the acoustic signal is used to extract prominence of the top word hypothesis in each slot using the boundaries extracted from CN.

Some of the ASR output scores are also inevitable parts of the island-gap classifier to measure how confident ASR is about the word hypotheses.The following features are used from the confusion networks: normalized posterior probability (*post*), normalized likelihood score (acoustic score per frame, (*as*), language model score (*ls*), To measure the uncertainty within a time slot, we also used entropy (*entropy*) of the probability distribution of the words within the time interval and competing posterior probability (*comp_post*: the ratio between posterior probability of the first-best and the second-best hypotheses within a time interval).

During LM scoring, when there is no entry in the LM for the higher order statistics of a word sequence, the speech recognizer uses the available lower order statistics. Hence, this is an evidence that shows how reliable the LM score is. The LM back-off values (NG) are printed at the lattice output; i.e., if LM score is as a result of 3-gram statistics then NG=2, if it is as a result of unigram statistics then NG=0. We used the following LM back-off related parameters: value of NG for the top hypothesis word in the time interval (*NG*); distance to max (*max_NG*) and min (*min_NG*): the difference between the value of NG that belongs to the top hypothesis and the maximum/minimum NG over all the words in a time slot; range of NG (*range_NG*: the difference between the maximum NG and minimum NG over all the words in a time slot).

4. Experiments and Results

The Boston University Radio News Corpus (BU-RNC) which consists of 3 hours read speech from 6 speakers (3 female, 3 male) was used in the experiments [14]. The database has manually labeled pitch accent tags which are only used during training for the prominence detection (using only training set). After eliminating story repetitions from the same speaker, the remaining data was split into five folds each with 50% train (14.5K words), 30% development (8.6K words), and 20% test (5.9K) sets. The Hidden Markov Model Toolkit (HTK) is used for the baseline experiments [15]. We adapted context-dependent triphone acoustic models trained from the WSJ and TIMIT tasks with data from the training partitions of the BU-RNC using the MAP and MLLR algorithms. The adapted acoustic models were gender specific (not speaker dependent). The 39-dimensional standard MFCC features are used as acoustic features. A standard back-off trigram language model with Kneser-Ney smoothing trained with the data from the CSR project was used. The language model vocabulary contained about 20K words. The OOV rate on the development and test sets were 3.8% and 3.7% respectively. The development set was used for tuning the scale parameters. The ASR 1-best hypothesis output is used as the baseline result. Also, lattices created using HTK are transformed to word confusion networks using the SRILM toolkit. The Wilcoxon signed rank test is used to report the confidence level in terms of significance values (p-values) whenever we make comparisons.

The development and train sets are used for training island-gap classifier. A 3-layer neural network is used as a classifier for island-gap detection. The neural network had D inputs, $(D + N)/2$ hidden nodes and N output nodes, where D is the length of feature vector, and $N = 2$ since this is a two class problem. Then, the information gain criteria is used to select the features using a forward algorithm; i.e., more features are added until the classifier accuracy starts to decrease. In Table 1, the features are ranked based on the information criteria. Among the features, entropy and prominence were the most and the least informative features, respectively, about the island-gap classes. This indicates that even though prominence is an important cue for humans, since previous stages of ASR ignore this cue, prominence has no information about the reliability of the created word hypotheses. The number of selected features that gives the highest accuracy for each fold varied from nineteen to twenty two. In Table 2, the island-gap detection results are presented. The chance level for the development and test sets were 79.5% and 78.4% respectively. With the proposed features, we achieved an overall 84.7% and 83.7% accuracy on

Table 1: Island-Gap Detection Features Ranked by Information

Rank	Feature	Rank	Feature
1	*entropy*	12	*as*
2	*posterior*	13	*n_ph*
3	*comp_post*	14	*min_dist*
4	*ls*	15	*f_syl*
5	*ncomp*	16	*ph_dur*
6	*unigram*	17	*range_NG*
7	*NG*	18	*n_syl*
8	*max_dist*	19	*duration*
9	*max_NG*	20	*min_NG*
10	*comp_dist*	21	*l_sylv*
11	*n_dist*	22	*prom*

Table 2: Island-Gap Detection Results

System	Overall Acc		Island Acc		Gap Acc	
	Dev.	Test	Dev.	Test	Dev.	Test
Predic.	84.7	83.7	93.9	94.5	48.9	44.9

the development and test sets, which are well above the chance level. The results are significantly higher than the previously reported results on island-gap detection; 63.47% island-gap classification accuracy was obtained in [9], however they used a different database.

In Table 3, the baseline results using the standard ASR 1-best output are presented. 18.6% and 18.4% word error rates (WER) were obtained on the development and test sets, respectively. We also tried rescoring the word confusion networks with the LM without using island-gap information, and this provides only 0.1% improvement over the baseline in both development and test sets as shown in Table 3.

In Table 4, the results obtained with using ASD with hard decision are presented. In the oracle experiments, where it is assumed that all the islands and gaps are known perfectly, the WER is reduced to 15.9% and 15.8%, providing 2.7% and 2.6% absolute improvements over the baseline in development and test sets, respectively. When the predicted island-gap information is used from the classifier, we obtained 18.2% and 18.0% WER on the development and test sets, respectively. This provides 0.4% (2.2% relative) improvement over the baseline in both sets, which is significant at $p \leq 0.001$.

In Table 5, the results obtained with using ASD with soft decision are presented. In the oracle experiments, the WER is further reduced to 15.4% and 15.5%, providing 3.2% and 2.9% absolute improvements over the baseline in development and test sets, respectively. The improvement over the hard decision oracle results is attributed to the pruning of the top word hypothesis, which is wrong, for gaps. When the automatically detected island-gap information is used from the classifier, we obtained 18.2% and 18.0% WER on the development and test sets, respectively. Similar to hard decision, this provides 0.4% (2.2% relative) improvement over the baseline in both sets, and the improvement is significant at $p \leq 0.001$. We observed that ASD with soft decision performed better than using hard decision with the automatically detected islands, however the improvement was not significant enough.

5. Conclusion

We presented an attention shift decoding method inspired by human speech recognition. In contrast to traditional ASR systems, ASD decodes speech inconsecutively using reliability criteria; the gaps (unreliable speech regions) are decoded with the evidence of islands (reliable speech regions). In the experiments with oracle information, ASD provides significant improvement (2.9% absolute) over the baseline ASR results confirming the promise of the method. At the heart of the ASD method is

Table 3: The Baseline ASR

System	Dev. WER	Test WER
Baseline	18.6	18.4
CN Rescoring	18.5	18.3

Table 4: The Results Using ASD with Hard Decision

System	WER		Improvement		Relative Improv.	
	Dev.	Test	Dev.	Test	Dev.	Test
Oracle	15.9	15.8	2.7	2.6	14.5%	14.1%
Predic	18.2	18.0	0.4	0.4	2.2%	2.2%

Table 5: The Results Using ASD with Soft Decision

System	WER		Improvement		Relative Improv.	
	Dev.	Test	Dev.	Test	Dev.	Test
Oracle	15.4	15.5	3.2	2.9	17.2%	15.8%
Predic	18.2	18.0	0.4	0.4	2.2%	2.2%

the automatic island-gap detection. Hence, we proposed a new feature set for island-gap detection and obtained 83.7% accuracy which is significantly higher than previously reported results. To cope with the imperfect nature of island-gap classification, we proposed a new ASD algorithm using soft decision rather than hard decision. The ASD with soft decision provided 2.2% relative improvement over the baseline which is significant ($p \leq 0.001$). As part of future work, we plan to explore more features to improve the island-gap detection accuracy.

6. References

[1] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Front. Biosci.*, vol. 5, pp. d202–212, 2000.

[2] I. Pollack and J. M. Pickett, "Intelligibility of Excerpts from Conversation," *The Journal of the Acoustical Society of America*, vol. 35, 1963.

[3] M. C. Silveri and A. Cappa, "Segregation of the neural correlates of language and phonological short-term memory." *Cortex*, vol. 39, no. 4-5, 2003.

[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 Computing Surveys (CSUR)*, 1980.

[5] T. Kawabata and K. Shikano., "Island-driven continuous speech recognizer using phone-based hmm word spotting," in *Proc. of ICASSP*, Glasgow, UK, May 1989.

[6] S. H. Lee, H. K. Lee, and J. H. Kim, "On-line Cursive Script Recognition Using an Island-Driven Search Technique," in *Int. Conf. on Document Analysis and Recog.*, vol. 2, 1995.

[7] J. F. Pitrelli, J. Subrahmonia, and B. Maison, "Toward island-of-reliability-driven very-large-vocabulary on-line handwriting recognition using character confidence scoring," in *Proc. of ICASSP*, vol. 3, 2001.

[8] A. Corazza, R. D. Mori, R. Gretter, and G. Satta, "Stochastic context-free grammars for island-driven probabilistic parsing," in *Proc. of IWPT*, Cancun, Mexico, 1991.

[9] R. Kumaran, J. Bilmes, and K. Kirchhoff, "Attention shift decoding for conversational speech recognition," in *Proc. of Interspeech*, Antwerp, Belgium, August 2007.

[10] 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*, 2000.

[11] J. M. McQueen, "Eight questions about spoken-word recognition," in *The Oxford handbook of psycholinguistics*, G. Gaskell, Ed., 2007.

[12] B. Fisher, "Syllabification software," http://www.itl.nist.gov/iad/mig//tools/, National Institute of Standards and Technology, 1997, .

[13] O. Kalinli and S. Narayanan, "Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information," *IEEE Trans. on Speech, Audio and Lang. Proc.*, 2009.

[14] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, *The Boston University Radio Corpus*, 1995.

[15] S. Young. (1989) Hidden markov model toolkit (HTK). [Online]. Available: http://htk.eng.cam.ac.uk/