A Study on Detection Based Automatic Speech Recognition

Chengyuan Ma, Yu Tsao and Chin-Hui Lee

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
{cyma, yutsao, chl}@ece.gatech.edu

Abstract

We propose a new approach to automatic speech recognition based on word detection and knowledge-based verification. Given an utterance, we first design a collection of word detectors, one for each lexical item in the vocabulary. Some pruning strategies are used to eliminate obviously unlikely words. Then these detected words are combined into word strings. The proposed approach is different from the conventional maximum a posteriori decoding method, and it is a critical component in building a bottom-up, detection-based speech recognition system in which knowledge in acoustic, speech and language can easily be incorporated in pruning unlikely word hypotheses and rescoring. The proposed approach was evaluated on a connected digit task using phone models trained from the TIMIT corpus. When compared with state-of-the-art connected digit recognition algorithms, we found the proposed detection based framework works well even no digit samples were used for training the detectors and recognizers. Other knowledge based constraints, such as place of articulation detectors, can be easily incorporated into this detection-based approach to improve the robustness and performance of the overall system.

1. Introduction

Researches on automatic speech recognition (ASR) have achieved dramatic progress and great success in the last several decades, due to the extensive use of statistical learning techniques, more and more speech and language data collections. In recent years, more improvements have been obtained in the field of speech and language modeling. The developments in detailed modeling and parameters sharing result in a higher acoustic resolution. The adaptive learning techniques can improve the model effectiveness when only sparse training data are available. And the criterion of discriminative learning algorithms is consistent with the speech recognition objective. So it can improve the model separation and recognition accuracy [1]. However, some challenging problems still exist within the prevailing ASR framework. One of them is the robustness in adverse conditions. The acoustic mismatch between the training and testing will cause the ASR performance to drop a lot. Meanwhile, linguistic mismatches, such as out of vocabulary and out of grammar, will bring more troubles to the ASR system. One reason for these limitations is that current ASR framework is a top-down, data-driven black box. That is, it provides very little diagnostic information for error correction and further improvement. Most ASR robustness issues are caused by ignoring the detail knowledge in acoustics, speech, language and their interactions. Some attempts were conducted to find robust distinctive feature which are invariant to speaker and speaking environments [2] [3]. Meanwhile, many knowledge supplemental modeling techniques have been investigated to incorporate available knowledge sources into state-of-the-art HMM based ASR system. But it's difficult to incorporate many knowledge sources into a single search network as required by the MAP decoding paradigm.

When compared with human speech recognition (HSR), state-of-the-art ASR systems usually have a much larger error rate even in clean environment. There is strong evidence that human speech recognition starts at a bottom-up analysis [4]. Then multiple knowledge sources are integrated into the recognition process. To bridge the performance gap between the state-of-the-art ASR system and HSR, a new detection-based, knowledge rich speech recognition paradigm has been proposed [5]. This new paradigm implies a new approach to solving the robustness problem and also can take advantages of the many other researches in phonetics, acoustics and linguistics. In this new paradigm, conventional data-driven statistical learning algorithms for ASR can be further extended by incorporating many knowledge sources. The detection-based ASR paradigm is very flexible in integrating many different kinds of knowledge sources. Because knowledge about the speech is explicitly built into the ASR system , the error correction and improvement can be made in a directed and meaningful manner.

In this paper, we demonstrate one implementation of this detection-based, knowledge rich ASR framework. Our proposed framework of the detection-based ASR is shown in Fig. 1. It consists of three parts: (1) word detectors design (2) knowledge guided word hypothesis verification and false alarm pruning (3) combining word hypothesis into word string.

When compared with state-of-the-art connected digit recognition algorithms we found the proposed detection based framework works well even no digit samples were used for training the detectors and recognizers. Other knowledge based constraints, such as place of articulation detectors, can be easily incorporated into this detection-based approach to improve the robustness and performance of the overall system.

Figure 1: *Framework of Detection-based ASR System.*

2. Word detector design

Many existing techniques (e.g., ANN, SVM and HMM) and many knowledge sources can be used for designing detectors at different levels, e.g., word level, sub-word level and attribute level. [10]. For connected digit recognition system, all the detectors are on the word level. We have a separate detector for each lexical item in the vocabulary. One of the basic principles for designing detectors is detect as many candidates as possible to avoid candidates missing. That is, we expect to have many false alarms while keep the missing rate as close to zero as possible. In this implementation, HMM modeling techniques are used for detector design. For each digit, either a whole-word model or a set of monophone models are trained from the training set. The key issue for HMM based detector design is how to choose an appropriate grammar network. A simple and intuitive grammar network of word detector is shown in Fig. 2. For each target word, it will compete with its corresponding anti-model and a silence model when decoding. The drawback of this design is that it will result in many missing errors.

Figure 2: *Simple Network of Digit Detector.*

A more complicated and elaborate network for word detector is shown in Fig. 3. Now for each target word, we introduce its cohort models and a silence model as the filler to absorb all the other events except for target word. With this network, less missing will occur. This is a very general detector design. One practical issue is how to select the cohorts for each target word. As a extreme example, for each target digit, we can choose all the other digits as its cohorts.

Figure 3: *General Network of Digit Detector.*

Fig. 4 shows an example of output of 11 digit detectors. The first and second panel are the test utterance $31o2$ and its spectrogram. The following 11 panels are the output of 11 detectors. For each panel, the segments above the horizon are the detected target segments and its magnitude is the score metric. For example, the last panel have three segments above the horizon. It means that the 'oh' detector tells us these segments are digit 'oh'. Actually, only the second segment is really a digit 'oh'. The first and the third one are false alarms.

Figure 4: *Hypothesis Generated by 11 Detectors.*

3. Word verification and pruning

It's obvious that these detectors generate a lot of false alarms just as we expect. To improve the recognition performance and reduce the computational complexity of the recognition process, it's desirable to verify these digit hypothesis and prune some of the false alarms. Word verification is essentially a statistical hypothesis testing problem [6] [7]. The likelihood ratio or generalized likelihood ratio is a good testing statistic for verification. One practical issue is to determine the threshold to accept the detectors output or reject it. Knowledge guided hypothesis verification and pruning is at the core of the detection-based ASR paradigm. All kinds of knowledge sources available from the acoustic, phonetic and linguistic research can be exploited for false alarm elimination. In the following, three pruning strategies will be presented.

3.1. Temporal information based pruning

For example, phoneme dependent duration constraints is one simple pruning strategy. The duration constraints can be used to eliminate those short segment in the detection result. The statistics of phoneme duration can be obtained from the training set. For example, the duration of one (/w/-/ah/-/n/) should be greater than 150 ms.

3.2. Attributes model based pruning

Another method is to use the manner attributes and place attributes model to generate the attributes sequence for each detected segment. Each manner attributes was modeled with a HMM. Then for each detected segments, it can be decoded as a sequence of manner attributes. If correctly decoded, each digit has its own attribute sequence pattern. Any obvious violations to the desired pattern can easily pruned out by some simple rules. For example, among all the output of detector "one", some of them are actually from "nine". So we can prune out those segments whose manner attributes sequence doesn't contain glides. This kind of model based pruning techniques have shown their effectiveness in our evaluation experiments.

3.3. Feature based pruning

The model based pruning is easy to be implemented and used. However, we still need to train these manner attributes model from some training set. Inevitably, the robustness problem still exists. So it's desirable to have some robust pruning strategies. Feature based pruning is one of them. For example, from the research of acoustics, we know that the energy of a nasal sound /n/ is concentrated on the low frequency region (below 400 HZ), while the fricative sound /f/ has a relatively flat spectrum. So this low frequency energy ratio feature is very useful and robust in distinguishing the nasal and fricative sound. Also the formants position of vowels and other spectral features can be used to distinguish certain pair of sounds [8].

4. Hypotheses combination

After the hypothesis verification and false alarm pruning, how can we combine these hypotheses from all detectors into a word string efficiently and accurately and what kind of criterion should be used to find the best word string?

4.1. Hypothesis lattice conversion

The weighted directed graph (WDG) is one of the methods that can be used to combine the detector output into a digit string. The hypotheses combination can be formulated as a search problem on a weighted directed graph. A weighted directed graph G is a pair (V, E) , where V is a set of vertices, and E is a set of edges between the ordered vertices $E = \{(u, v)|u, v \in V\}$. Meanwhile, there is a weight $W_{u,v}$ associated with each edge.

The following procedure can be used to convert the hypothesis lattice into a directed graph.

1. Constructing the nodes set V. V consists of all the detected digit boundary. For instance, one detector detected a segment (T_a, T_b) as a target hypothesis. So both T_a and T_b will be elements of V.

2. Rank all the detected boundary in a time line and add a edge for each pair of adjacent nodes in the graph in order to guarantee the existence of a path from start node to end node.

Figure 5: *Weighted Directed Graph.*

3. For each detected segment, adding a edge from its start node to its end node.
4. Add reversal edge to those nodes which are very close to each other. (e.g. within 20 ms) or merge these nodes into one node, because the potential overlap in the detected boundaries.

4.2. Search in the weighted directed graph

Given the constructed directed graph, how can we assign the appropriate weight to each edge? Obviously, the weight we choose should be consistent with our search criterion. For example, when the search criterion is the maximum likelihood criterion, the log-likelihood should be used as the weight. Of course, we can put other score metrics to each edge under certain criterion.

Finding the best path in a WDG is a well studied task in computer science and operational research. So finding the best matched string over the detector output lattice is equivalent to find a path with the maximal weight. The well-known Dijkstra's algorithm can be used to find the best matched path. To further improve the recognition performance by rescoring with other detectors' results, the KSP (K-Shortest Path) algorithm [9] can be used to find the k-best digit strings. Fig. 5 is the WDG converted from Fig. 4. Each node in the graph is a detected digit boundary. The number in the node is the time (in 10 ms). Each edge represent a detected digit or a silence edge. The number beside each edge is the frame average log-likelihood. And the red edges are the best path we got for the utterance $31o2$.

5. Experiment setup and result analysis

All the evaluation experiments are conducted on the TIDIGITS corpus [12]. The TIDIGITS corpus vocabulary is made of 11 digits, one to nine, plus oh and zero. The training set has 8623 digit strings and the test set has 8700 digit strings. A conventional procedure is used for front-end processing. 12-dimensional MFCC and the log-scaled energy were extracted for each 10-ms frame. Their first and second order derivatives are also computed for each frame. To conduct cross corpus evaluation and reduce the channel effects, every dimension of the feature vector has been normalized with zero-mean and unit variance.

5.1. Whole word model in matched condition

In this experiment, the training set from the TIDIGITS corpus are used to train the whole-word HMM model for each digit. Each HMM has 12 states and use a simple left-to-right topology without state-skip. A state-of-the-art HMM based ASR system and a detection-based ASR system are built for comparison. The conventional HMM based ASR got a word accuracy about 99.52% and the detection-based ASR got 99.18%. So in the matched acoustic condition, the detection-based system can get comparable results as the conventional ASR system.

5.2. Monophone model in mismatched condition

This experiment is a simulation of real ASR scenario. There exist acoustic mismatch between the training set and testing set. TIMIT [11] was used for mono-phone model training while the TIDIGITS was used for testing. Each mono-phone model is a 3-state left-to-right HMM. A conventional ASR system and a detection-based ASR system was built for this experiments. The experiment results are shown in Table 1.

The word accuracy of the conventional ASR system is 95.46%. And for the detection-based ASR system, the word accuracy is 93.63%. It's clear that the detection-based system has much more substitution and insertion errors.

When we took a close look at the recognition results of the detection-based ASR system, we found too many short segments were detected and recognized as a word. So the phoneme-dependent duration constraints can be imposed on the detection results. For example, the duration of word one should be at least 150 ms. Otherwise, it could be a false alarm. After pruning with the duration constraints, the performance of the detection based ASR system was improved to 94.97%. The insertion errors has been reduced from 791 to 351, while the deletion errors increase from 167 to 227.

Some confusion pairs are very significant in the word confusion matrix. For example, five/nine (ground-truth/recognized result), five/four, one/nine, eight/three, seven/five, four/oh, etc. Some of these substitution errors can be easily fixed by abovementioned manner model based pruning. For example, for five/four confusion, if the manner sequence for the detected 'four' segment doesn't contain the glides, it will be pruned. With such a simple rule, the substitution error of five/four is reduced from 51 to 7. Other similar rules can be used for false alarm elimination on the detection result of one and nine. The overall performance after manner model based pruning is 95.50%. It's the same as the performance of the conventional ASR system. We can see that the substitution errors were reduced from 860 to 720 and the insertion error were reduced from 351 to 305.

Feature based pruning is more meaningful and robust. The spectral features of nasal and fricative can be used in five/nine confusion pair. The substitution errors of five/nine were reduced from 41 to 15 by using the low frequency energy ratio and voicing detector. As for the eight/three confusion pair, the spectrum before the /iy/ in three and /ey/ in eight are different. With voicing detector and high-frequency energy concentration, we can reduce the substitution of eight/three from 56 to 24. Now the overall performance is improved to 95.76%. The substitution errors have been further reduced (from 720 to 631), while the deletion errors increased a little (from 262 to 280).

Although this result is only slightly better than the result of convention ASR system, this kind of feature based pruning is very promising. If we use the training data from TIDIGITS to get better manner models for pruning, the performance will be 96.48%. It's much better than the result of conventional state-of-the-art ASR system. It shows that even if the acoustic model for detector design is not perfect, we can still have very good recognition performance by word detection and appropriate pruning.

6. Summary and future work

In this paper, we demonstrated one implementation of the detection-based, knowledge rich speech recognition paradigm. Our experiment results show that by explicitly incorporating our knowledge about the speech and language into our detector design and pruning strategy, the performance of the detection-based ASR system can be improved step by step in a meaningful and directed manner. It's also noted that the performance in the proposed sys-

Table 1: *ASR result.*

	Del.	Sub.	Ins.	Word Acc. (%)
Detection W/O Pruning	167	864	791	93.63
W/ Duration Pruning	227	860	351	94.97
W/ Manner Pruning	262	720	305	95.50
W/ Feature Pruning	280	631	301	95.76
Upper Bound	294	415	296	96.48
Conventional ASR	469	617	211	95.46

tem is additive. That is, a better module for a feature will not produce poorer performance for the individual module and overall performance. The word verification and pruning strategies mentioned in this paper are still faraway from being perfect. We are expecting more reliable knowledge sources can be detected. In future studies, more knowledge sources will be incorporated into the framework for hypothesis pruning. Also some post-processing can be done on the N-best candidates. We are more interested in investigating the detection-based ASR system for LVCSR tasks.

7. References

[1] Lee, C.-H., "On Automatic Speech Recognition at the Dawn of the 21st Century," *IEICE Trans. Inf. & Syst.*, pp. 377–396, 2003.

[2] Liu, S. A., "Landmark Detection for Distinctive Feature-based Speech Recognition," *JASA*, pp. 3417–3430, 1996.

[3] Juneja, A. and Espy-Wilson, C., "Segmentation of Continuous Speech Using Acoustic-Phonetic Parameters and Statistical Learning," *Proc. ICONIP*, vol. 2, pp. 726–730, 2002.

[4] Allen, J. B., "How Do Humans Process and Recognize Speech?" *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[5] Lee, C.-H., "From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition," in *Proc. InterSpeech*, 2004.

[6] Sukkar, R. A. and Lee, C.-H., "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 6, pp. 420–429, 1996.

[7] Kawahara, T., Lee, C.-H., and Juang, B.-H., "Flexible Speech Understanding Based on Combined Key-phrase Detection and Verification," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 6, pp. 558–568, 1998.

[8] Rabiner, L. R. and Schafer, R. W., *Digital Processing of Speech Signals*, 1993, Prentice Hall

[9] Epstein, D., "Finding the K-Shortest Paths," in *SIAM J. Computing*, pp. 652–673, 1998.

[10] Li, J., Tsao, Y. and Lee, C.-H., " A Study on Knowledge Source Integration for Candidate Rescoring in Automatic Speech Recognition," in *Proc. ICASSP*, 2005.

[11] Garofalo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," Tech. Rep., U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[12] Leonard, R. G., "A Database for Speaker-Independent Digit Recognition," in *Proc. ICASSP*, 1984.