EMERGENT: JOURNAL OF EDUCATIONAL DISCOVERIES AND LIFELONG LEARNING (EJEDL)
ISSN: 2776-0995 Volume 2, Issue 11, Nov, 2021

DESIGN AN AUTOMATIC SPEECH VERIFICATION BASED ON WORD DETECTION

Arshad B. Salih
Northern Technical University, Technical Institute Kirkuk
arshad.b.salih@ntu.edu.iq

Abstract
Based on word identification and knowledge-based verification, we offer a new technique to automated speech recognition (ASR). Assuming we know the vocabulary, we begin by creating word detectors for each word in the utterance. Pruning techniques are employed to weed out improbable word candidates. These words are then grouped together to form word strings. To construct a bottom-up, detection-based voice recognition system that incorporates knowledge of acoustics, speech, and language into pruning and rescoring, the suggested strategy differs from the typical maximal a posteriori decoding method. Using phone models learned from the TIMIT corpus, the suggested method was tested on a connected digit task. Even though no digit samples were provided to train the detectors and recognizers, the suggested detection-based framework performed well when compared to current linked digit recognition methods. This detection-based technique can incorporate other knowledge-based limitations, such as the manner and location of articulation detectors, to enhance the overall system's robustness and performance.

Keywords: speech recognition, detection-based.

Introduction

Research on automatic speech recognition (ASR) has witnessed dramatic progress and great success in the last several decades. More improvements have been obtained in the field of speech and language modeling due to the extensive use of statistical learning techniques, more and more speech and language data collections. 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 events will cause misrecognition. 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. Furthermore, ASR robustness issues are often caused by ignoring the detail knowledge in acoustics, speech, language and their interactions. One way to incorporate knowledge sources into ASR system designs is through bottom-up detection of fundamental speech unit followed by knowledge integration [1]. 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 hidden Markov model (HMM) based ASR system. But it's difficult to incorporate many knowledge sources into a single search network as required by the maximum a posteriori (MAP) decoding paradigm. Whencompared 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 realize such a knowledge-driven ASR framework, a new detection based, knowledge-rich speech recognition paradigm has been proposed [5]. It implies a new approach to solving the robustness problem and also takes advantages of the rich literatures in phonetics, acoustics and linguistics. Conventional data-driven statistical learning algorithms for ASR can also be further extended by incorporating diverse knowledge sources. The detection-based ASR paradigm is 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 study, we demonstrate one implementation of this detection-based, knowledge-rich ASR framework. Our proposed framework of the detection-based ASR is shown in Figure. 1. It consists of three parts: (1) word detectors design; (2) knowledge

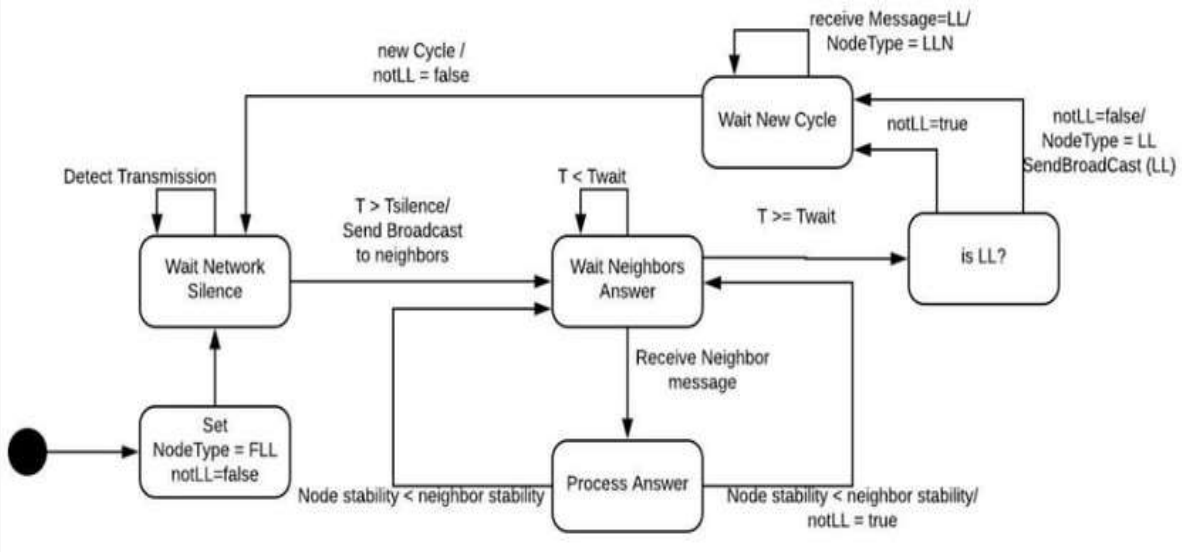

Figure 1. ASR framework

2-Word Detector Design

Many existing techniques (e.g., artificial neural network (ANN), support vector machine (SVM) and HMM) and many knowledge sources can be used for designing detectors at different stages, e.g., word, sub-word and attribute levels. 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 to detect as many candidates as possible to avoid candidates missing. That is, we expect to have many false alarms while keeping the missed detection rate as close to zero as possible. In this implementation, HMM modeling techniques are used for detector design. For each digit, a set of monophony 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 example for detecting a word is shown in Figure. 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 missed detection errors.

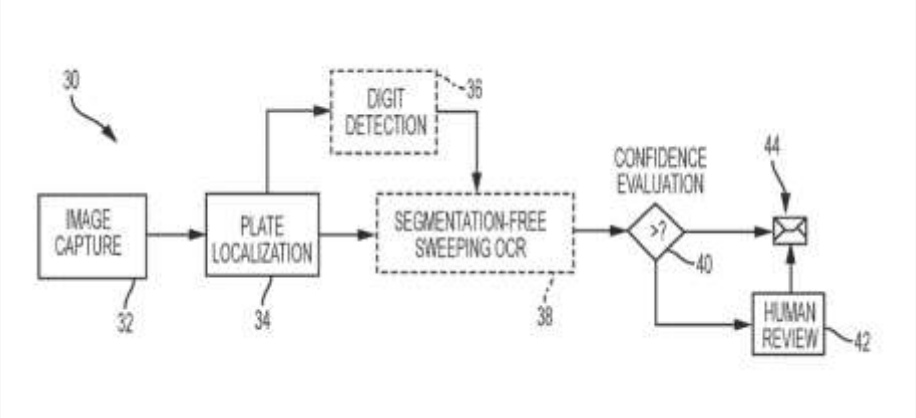

Figure 2. Digit detector system

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

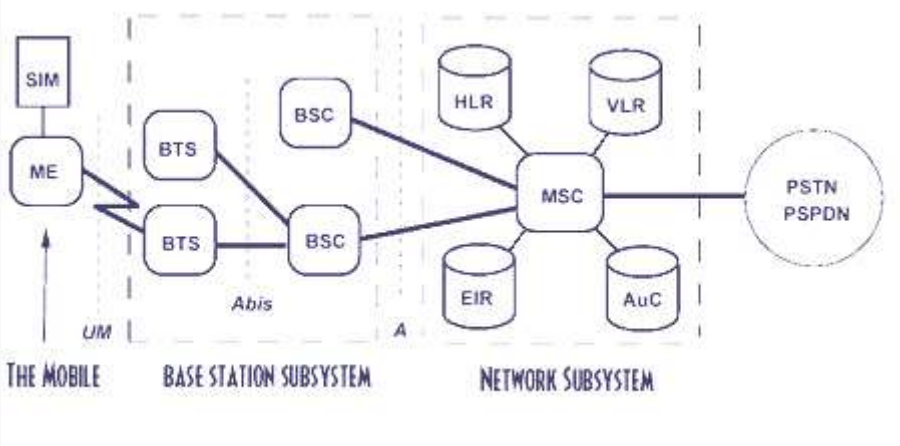

Figure 3: General Network of Digit Detector

EMERGENT: JOURNAL OF EDUCATIONAL DISCOVERIES AND LIFELONG LEARNING (EJEDL)
ISSN: 2776-0995 Volume 2, Issue 11, Nov, 2021

3. Word Verification and Pruning

It is obvious that these detectors generate many false alarms just as we expect. To improve the recognition performance and reduce.

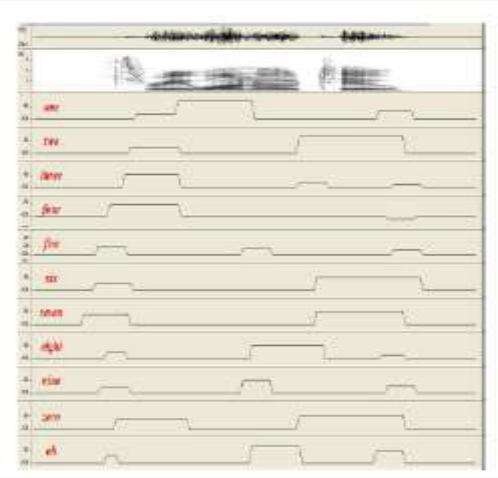

Figure 4: Hypotheses Generated by 11 Detectors

The computational complexity of the recognition process, it is desirable to verify these digit hypotheses and prune some of the false alarms. Word verification is formulated as a statistical hypothesis testing problem [6]. 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 detector outputs or reject them. Knowledge guided hypothesis verification and pruning is at the core of the detection-based ASR paradigm. Al knowledge sources available from acoustic, phonetic and linguistic research can be exploited for false alarm elimination. In the following, three pruning strategies will be presented. We expect more will come out from the community.

3.1. Temporal information based pruning

For example, phone dependent duration constraint is one simple pruning strategy. The duration constraints can be used to eliminate those short segments in the detection result. The statistics of phone duration can be obtained from the training set.

3.2. Attributes model based pruning

Another method is to use models of the manner and place attribute to generate the attribute sequence for each detected segment. Each manner attribute is modeled with a HMM. Then fo each detected segment, it can be decoded as a sequence of manner attributes. If correctly decoded, each word has its own attribute sequence pattern. Any obvious deviation from the desired pattern.

Can be pruned by some rules. For example, among all the output of detector "one", some of them are actually from speech for "nine". So we can prune those segments whose manner attribute sequence doesn't contain glides. This kind of model based pruning techniques have shown their effectiveness in our evaluation experiments.

3.3. Signal Based Pruning

Model based pruning can easily be implemented and used. However, we still need to train these manner attribute models from some training set. Inevitably, the robustness problem still exists. So it's desirable to have some robust pruning strategies. Signal feature based pruning is one of them. For example, from research in acoustics, we know that the energy of a nasal sound /n/ is often concentrated on the low frequency region (below 400 HZ), while the fricative sound /f/ has a relatively flat spectrum and energy. Distribution in high frequency region. So the percentage of low frequency energy in the total energy is 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.

4. Experiment Setup and Result Analysis

All the evaluation experiments are conducted on the TIDIGITS corpus [9]. The digit 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 element of the feature vector has been normalized with zero-mean and unit variance.

4.1. 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 gave a word error rate about 0.48% and the detection-based ASR was slightly worse at 0.73%. Therefore, in the matched acoustic condition, the detection-based system can get comparable results as the conventional ASR system.

4.2. Monophone Model in Mismatched Condition

Now we simulate a real ASR scenario. We purposely introduced a mismatched condition to illustrate the benefits of incorporating knowledge into the detection based ASR system. TIMIT [8] was used for mono-phone model training while the TIDIGITS was

down-sampled from 20 KHz to 16 KHz and used for testing. Each mono-phone model is a 3-state left-to-right HMM. A conventional Viterbi-based ASR system and a detection-based ASR system were built for the experiment. The deletion, substitution and insertion errors of step-by-step knowledge-based pruning are shown in Table 1. The word error rate of the conventional ASR system is 4.54%. And for the detection-based ASR system without pruning, it is 6.37%. It's clear that the detection-based system has much more substitution and insertion errors.

Duration Pruning: 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 words. So the phone-dependent duration constraints can be imposed on the detection results. After pruning with the duration constraints, the word error rate of the detection based ASR system was reduced to 5.03%. The insertion errors were reduced from 791 to 351, while the deletion errors increase from 167 to 227.

Manner Pruning: We also observed that 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 reduced by manner model based pruning The rules used for pruning can be learned from some development data by decision tree. The manner sequence pattern pruning method can generally be used to prune those clear confusions. The overall performance after manner model based pruning is 6.23%. We can see that the substitution errors were reduced from 864 to 520 and the insertion error were reduced from 354 to 304.

5. Conclusion

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 systematically in a meaningful and directed manner. It's also noted that the performance improvement in the proposed system is additive. That is, a better module for a feature will not produce as much poorer result for the individual module and overall performance. The word verification and pruning strategies mentioned in this paper are still far away from being perfect. We are expecting more reliable knowledge sources to be detected. In future studies, more knowledge sources will be incorporated into the framework for hypothesis pruning. In addition, 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

REFERENCES

1. Al Nuaimi, M., Shuaib, K., Al Nuaimi, K. (2014). Clustering in WSN Using Node Ranking with Hybrid Nodes Duty-Cycle and Energy Threshold. IEEE 13th International Symposium on Network Computing and Applications, Cambridge, MA. (21-23 August)
2. Nithyakalyani, S., Gopinath, B. (2015). Analysis of Node Clustering Algorithms on Data Aggregation in Wireless Sensor Network. Journal of Scientific & Industrial Research, 5(72), 38-42.
3. Nithyakalyani, S., Gopinath, B. (2015). Analysis of Node Clustering Algorithms on Data Aggregation in Wireless Sensor Network. Journal of Scientific & Industrial Research, 5(72), 38-42.
4. Nithyakalyani, S., Gopinath, B. (2015). Analysis of Node Clustering Algorithms on Data Aggregation in Wireless Sensor Network. Journal of Scientific & Industrial Research, 5(72), 38-42.
5. Yuhui, J., Junping, H. (2008). A Time- based Cluster - Head Selection Algorithm for LEACH. IEEE Symposium on Computers and Communications, Marrakech. (6-9 Haziran)
6. Yuhui, J., Junping, H. (2008). A Time- based Cluster - Head Selection Algorithm for LEACH. IEEE Symposium on Computers and Communications, Marrakech. (6-9 Haziran)
7. Lindsey, S., Raghavendra, C. S. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems. IEEE Aerospace Conference Proceedings. 3(3)1125 - 1130.
8. Lindsey, S., Raghavendra, C. S. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems. IEEE Aerospace Conference Proceedings. 3(3)1125 - 1130.
9. Lindsey, S., Raghavendra, C. S. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems. IEEE Aerospace Conference Proceedings. 3(3)1125 - 1130.
10. Lindsey, S., Raghavendra, C. S. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems. IEEE Aerospace Conference Proceedings. 3(3)1125 - 1130.
11. Lindsey, S., Raghavendra, C. S. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems. IEEE Aerospace Conference Proceedings. 3(3)1125 - 1130.
12. Lindsey, S., Raghavendra, C. S. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems. IEEE Aerospace Conference Proceedings. 3(3)1125 - 1130.
13. Lindsey, S., Raghavendra, C. S. (2002). PEGASIS: Power-Efficient Gathering in

Sensor Information Systems. IEEE Aerospace Conference Proceedings. 3(3)1125 - 1130.

14. Mahmood, Z., Nasret, A., Awed, A., Design of New Multiband Slot Antennas for Wi-Fi Devices, (2019) International Journal on Communications Antenna and Propagation (IRECAP), 9 (5), pp. 334-342.
15. Z. S. Mahmood, A. N. N. Coran and A. Y. Aewayd, "The Impact of Relay Node Deployment In Vehicle Ad Hoc Network: Reachability Enhancement Approach," 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-3, doi: 10.1109/GCAT47503.2019.8978445.
16. Mahmood, Z. S., Coran, A. N. N., & Kamal, A. E. (2018). Dynamic approach for spectrum sharing in cognitive radio. International Journal of Engineering & Technology, 7(4), 5408-5411.
17. Nasret, A., & Mahmood, Z. (2019). Optimization and integration of rfid navigation system by using different location algorithms. International Review of Electrical Engineering (IREE), 14(4).
18. Mahmood, Z. S., Nasret, A. N., & Mahmood, O. T. (2021, October). Separately excited DC motor speed using ANN neural network. In AIP Conference Proceedings (Vol. 2404, No. 1, p. 080012). AIP Publishing LLC.