Distributed Listening: A Parallel Processing Approach to Automatic Speech Recognition

Yolanda McMillian
3101 Shelby Center
Auburn University
Auburn, AL 36849-5347, USA
mcmilym@auburn.edu

Juan E. Gilbert
3101 Shelby Center
Auburn University
Auburn, AL 36849-5347, USA
gilbert@auburn.edu

Abstract

While speech recognition systems have come a long way in the last thirty years, there is still room for improvement. Although readily available, these systems are sometimes inaccurate and insufficient. The research presented here outlines a technique called Distributed Listening which demonstrates noticeable improvements to existing speech recognition methods. The Distributed Listening architecture introduces the idea of multiple, parallel, yet physically separate automatic speech recognizers called listeners. Distributed Listening also uses a piece of middleware called an interpreter. The interpreter resolves multiple interpretations using the Phrase Resolution Algorithm (PRA). These efforts work together to increase the accuracy of the transcription of spoken utterances.

1 Introduction

Research in the area of natural language processing has been on-going for over thirty years (Natural Language Software Registry, 2004; Jurafsky and Martin, 2000); however, there is still room for improvement with mainstream speech recognition systems (Deng, 2004). Distributed Listening will further research in this area. The concept is based around the idea of multiple speech input sources. Previous research activities involved a single microphone with multiple, separate recognizers that all yielded improvements in accuracy. Distributed Listening uses multiple, parallel speech recognizers, with each recognizer having its own input source (Gilbert, 2005). Each recognizer is a listener. Once input is collected from the listeners, one machine, the interpreter, processes all of the input (see figure 1). To process the input, a phrase resolution algorithm is used.

This approach is analogous to a crime scene with multiple witnesses (the listeners) and a detective (the interpreter) who pieces together the stories of the witnesses using his/her knowledge of crime scenes to form a hypothesis of the actual event. Each witness will have a portion of the story that is the same as the other witnesses. It is up to the detective to fill in the blanks. With Distributed Listening, the process is very similar. Each listener will have common recognition results and the interpreter will use the phrase resolution algorithm to resolve conflicts.

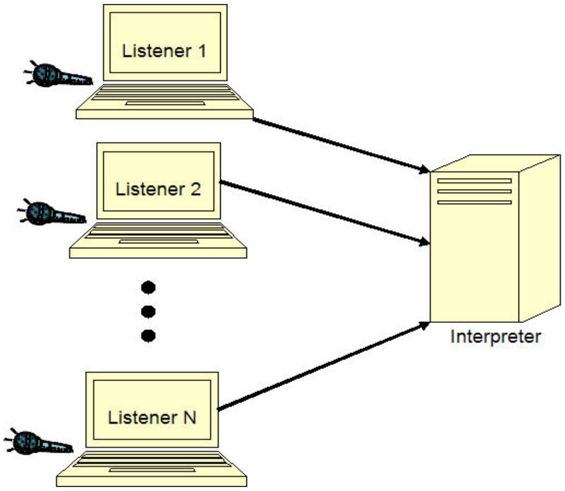

Figure 1. Distributed Listening Architecture

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 173–176,
Columbus, Ohio, USA, June 2008.

2 Background

Automatic speech recognition systems convert a speech signal into a sequence of words, usually based on the Hidden Markov Model (HMM), in which words are constructed from a sequence of states (Baum, 1972; Young et al., 1989; Young 1990; Furui, 2002).

There are several systems that used the HMM along with multiple speech recognizers in an effort to improve speech recognition, as discussed next.

2.1 Enhanced Majority Rules

Barry (et al., 1994) took three different Automatic Speech Recognition (ASR) systems, along with an Enhanced Majority Rules (EMR) software algorithm. Each of the three individual systems received the same input, performed speech recognition, and sent the result to the master system.

The EMR resolved inconsistencies by looking for agreement from the individual systems for the recognized word. If there was no majority agreement, the EMR looked to the second word for agreement before relying on the distance scores. This architecture produced better recognition accuracy than each of the individual systems.

While an improvement was made, the architecture can suffer from distorted input. Since each system receives the same input, if the input signal is not good, then all of the individual systems will receive bad input.

2.2 Virtual Intelligent Codriver

The Virtual Intelligent Codriver (VICO) project also used multiple ASR systems in parallel (Brutti et al., 2004; Cristoforetti et al., 2003). Each ASR received the same input and had its own language model. The resulting interpretations from each ASR are compared to each other using confidence scores. The interpretation with the highest recognition accuracy is selected. While the experiments resulted in noticeable improvements over the individual ASR systems, there are two shortcomings. First, if the input signal is distorted, then each recognizer will receive bad input. Secondly, if each recognizer contains a piece of the optimal interpretation, then this architecture falls short.

2.3 Recognized Output Voting Error Reduction

The Recognizer Output Voting Error Reduction (ROVER) system is a composite of multiple ASR systems that uses a voting process to reconcile differences in the individual ASR system outputs (Fiscus, 1997). Multiple interpretations are passed from each recognition engine to the alignment module. Once aligned, the voting module is called. The voting module scores each word within the alignment vertically and the words with the highest scores are chosen. On average, this composite ASR system produces a lower error rate than any of the individual systems, but suffers from order of combination and ties.

2.4 Modified ROVER

To solve the problem that results from the order of combination and ties of the original ROVER system, Schwenk proposed a modified ROVER system that used a dynamic programming algorithm built on language models (Schwenk and Gauvain, 2000). The modified ROVER system resulted in a reduction in the word error rates over the original ROVER system.

3 Distributed Listening

Distributed Listening builds on the architectures that use multiple speech recognizers and enhances it with the use of multiple input sources.

Distributed Listening is made of three significant parts: Listeners, an Interpreter, and a Phrase Resolution Algorithm.

3.1 Listeners

Distributed Listening uses multiple speech recognizers, working in parallel, to process the spoken input. Each recognizer is called a listener and is equipped with it's own input source. Each listener is a separate, physical computing device with its own memory, processor, and disk space. Each listener collects input from the user. The result of each listener is passed to the interpreter.

3.2 Interpreter

Once input is collected from the listeners, the input is passed to the interpreter. The interpreter will process all of the input collected from each listener as described next.

3.3 Phrase Resolution Algorithm

To resolve multiple interpretations from the listeners, the Phrase Resolution Algorithm (PRA) is used.

The underlying grammar of the PRA is based on an N-gram language model. An N-gram language model is used by the recognizer to predict word sequences. Distributed Listening uses an N-gram of size 1, also known as a unigram. The grammar consists of known utterances that can be made by the user.

The unigram grammar is stored in a phrase database. The grammar is organized according to individual words and phrases. Each phrase is placed in a table. The phrases are broken down into their individual words and placed in another table. The table of words keeps a count of the number of times each word appears in each phrase, resembling the unigram language model.

To determine the most likely spoken phrase, queries are made against the collection of individual words, also known as the complete word set. The queries try to identify matching phrase(s) based on specified words. The matching phrase(s) with the highest concentrations of words is returned by the query.

The word concentration is determined by comparing the length of the phrase with the number of matching words found in the complete word set. The concentration of the number of words found within each phrase is calculated using all interpretations from the listeners. The phrase(s) with the highest concentration of words is the most likely spoken phrase.

4 System Architecture

There are multiple models for Distributed Listening; Homogeneous, Heterogeneous, and Hybrid. The Homogeneous model uses the same grammar for each listener. Within the Heterogeneous model, each listener uses a different grammar. The Hybrid model contains a combination of the Homogenous and Heterogeneous models.

4.1 Homogeneous

In a homogenous Distributed Listening architecture, each listener has the same grammar or language model. Although all of the listeners are identical in capturing the input, this architecture allows for the different perspectives of the utterances to also be captured.

4.2 Heterogeneous

Heterogeneous architectures use different grammars or language models on each listener. Each listener has its own input source and recognizer and implies a distributed grammar/language model. This allows for flexibility as very large grammars and vocabularies can be distributed across several listeners.

4.3 Hybrid

The hybrid architecture is a homogenous architecture of heterogeneous Distributed Listening nodes, as shown in figure 2. This gives the embedded environment the ability to recognize multiple languages, as well as accommodate translations of inter-mixed spoken language.

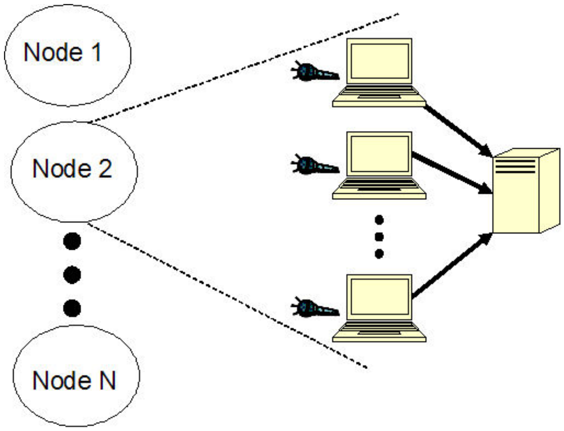

Figure 2. Hybrid Distributed Listening Architecture

5 Conclusion

The goal of Distributed Listening research is to take a unique approach in order to enhance the success of the traditional approaches to speech recognition. The approach of Distributed Listening directly mimics people. The psychology domain has shown that people use a form of Distributed Listening called Dichotic Listening, where people listen to two voices, one in each ear, at the same time (Bruder, 2004). Distributed Listening is a natural extension of Dichotic Listening, where computers are listening in the same manner as people. Distributed Listening is an attempt to enable computer systems to perform similar to humans while decreasing error rates.

Preliminary studies have shown a decrease in error rates. Early results indicate that Distributed Listening is a viable alternative to current speech recognition systems. Additional studies are being planned that will effectively test the Phrase Resolution Algorithm.

References

Barry, T., Solz, T., Reising, J. & Williamson, D. **The simultaneous use of three machine speech recognition systems to increase recognition accuracy**, In Proceedings of the IEEE 1994 National Aerospace and Electronics Conference, vol.2, pp. 667 - 671, 1994.

Baum, L.E. **An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process**. Inequalities 3, 1-8, 1972.

Bruder, G.E., Stewart, J.W., McGrath, P.J., Deliyannides, D., Quitkin, F.M. **Dichotic listening tests of functional brain asymmetry predict response to fluoxetine in depressed women and men**. Neuropsychopharmacology, 29(9), pp. 1752-1761, 2004.

Brutti, A., Coletti, P., Cristoforetti, L., Geutner, P., Giacomini, A., Gretter, R., et al. **Use of Multiple Speech Recognition Units in a In-car Assistance Systems**, chapter in "DSP for Vehicle and Mobile Systems", Kluwer Publishers, 2004.

Cristoforetti, L., Matassoni, M., Omologo, M. & Svaizer, P., **Use of parallel recognizers for robust in-car speech interaction**, In Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing [ICASSP 2003], Hong-Kong, 2003.

Deng, L. & Huang, X., **Challenges in adopting speech recognition**, Communications of the ACM, vol. 47, no. 1, pp. 69-75, January 2004.

Fiscus, J. G., **A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER)**. In IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347–354, 1997.

Furui, S., **Recent progress in spontaneous speech recognition and understanding**, In Proceedings of the IEEE Workshop on Multimedia Signal Processing, 2002.

Gilbert, J. E. (2005). **Distributed Listening Research.** *In Proceedings of AVIOS Speech Technology Track*, San Francisco, California, SpeechTEK West, pp. 1 – 10.

Jurafsky, D. & Martin, J., **Speech and Language Processing**, Prentice Hall, 2000.

Natural Language Software Registry, [Online]. Available: http://registry.dfki.de/, 2004.

Schwenk, H. & Gauvain, J., **Improved ROVER using Language Model Information**, In ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium, Paris, pp. 47–52, 2000.

Young, S.R., **Use of dialog, pragmatics and semantics to enhance speech recognition**, Speech Communication, vol. 9, pp. 551-564, 1990.

Young, S.R., Hauptmann, A.G. , Ward, W.H. , Smith, E.T. & Werner, P., **High level knowledge sources in usable speech recognition systems**, Communications of the ACM, vol. 31, no. 2, pp. 183-194, 1989.