Generalized Hierarchical Search in the ISIP ASR System

Bohumir Jelinek, Feng Zheng, Naveen Parihar, Jonathan Hamaker, and Joseph Picone

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, MS 39762 USA
email: jelinek@isip.mstate.edu

ABSTRACT

It has long been a goal of speech researchers to incorporate higher-level knowledge sources such as discourse, part of speech, and understanding constraints into the speech recognition problem. However, current speech recognition systems are highly tuned to N-gram, triphone-based recognition. Thus, researchers have been unable to exploit this knowledge without extensive modifications to the most complex portion of an ASR system - the decoder. In this paper, we describe a publicly-available, state-of-the-art decoder that employs a flexible and configurable multi-level search strategy capable of incorporating hierarchical knowledge sources with no changes to source code.

1. INTRODUCTION

State-of-the-art large vocabulary continuous speech recognition (LVCSR) systems use a search strategy that dynamically integrates a highly simplified form of the acoustic and linguistic constraints [1]. The problem of speech recognition is cast into a statistical framework where we try to find the most probable word sequence given observed acoustic signal, A [2]. We apply Bayes' rule to create the maximum likelihood formulation of the problem:

$$W^* = \arg\max_W \quad P(A|W)P(W). \tag{1}$$

P(A|W) is the probability that the observation sequence A occurred given that the word string W was spoken. This is typically provided by an acoustic model such as an HMM. P(W), is the prior probability of a sequence of words and is typically determined using an N-gram language model or a stochastic grammar. The search process combines these two probabilities for all possible word sequences and selects the one sequence with the maximum probability as the final hypothesis.

While simple and efficient, this formulation is not accurate. As shown in Figure 2, humans employ a variety of information sources for speech recognition. These include acoustic pattern recognition, linguistic pattern constraints, and syntactic and semantic pattern analysis. There has been much research studying each of these isolated information sources but we have not been able to integrate them into a single parsimonious framework. This is due, in large part, to the inflexibility of the basic search algorithms employed for speech recognition.

Incorporating such important features to produce a more robust recognition system often requires extensive algorithm modifications and software changes. This paper describes a generalized hierarchical speech recognition system. The graph search mechanisms therein provide extreme flexibility in defining the constraints of the system. The key feature of this system is the extension of the search structure to an unlimited number of hierarchical knowledge sources each with individually adjustable sets of parameters through the user of only a configuration file.

2. HIERARCHIAL SEARCH

Our generalized search algorithm is based upon a hierarchical graph-based level building approach. Each knowledge source in the hierarchy is assumed to be representable by knowledge sources. For example, the typical speech recognition framework can be easily fit into this paradigm. Words constitute the top level in the hierarchy. These can be decomposed into phonemes at the next level and likewise phonemes can be modeled by HMM state sequences at the bottom-most level. The novelty of our approach is that extend this to allow for an

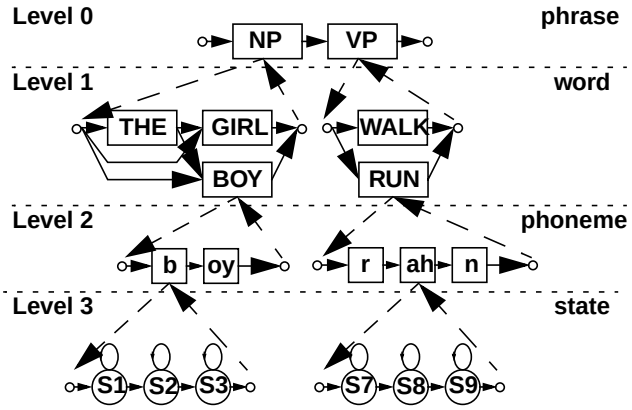

Figure 1: Integration of information from four knowledge sources at four different levels in a hierarchical structure.

Generalized Search	**Typical Search**
General Specification	N-gram based
Any number of independent levels	Word, phone and state level
Unlimited context	Context limited to one left and one right (triphone based)
Flexible Search Structure	Tuned to a specific task

Table 1: Comparison between the generalized search and a typical standard algorithm.

unlimited set of knowledge sources. This provides flexibility and extensibility so long as the knowledge sources can be stated in a hierarchical form — a requirement that we would argue is not very limiting.

For example, in Figure 1, we add part-of-speech knowledge to the hierarchy. The search is decomposed into the series of hierarchical levels, each with its own information source. The top level integrates syntactic information consisting of part-of-speech tags. This level is decomposed into words which specify a strict grammar. The words are decomposed into phonemes which may model pronunciation variability in the data. The pronunciations are specified in a general graph form so there is no restriction as to the number nor the form of the pronunciations. Finally, the phonemes are modeled by HMM states. Again, due to the generalized nature, the form of the HMMs is not limited (typical HMM systems require left-to-right systems with diagonal covariance Gaussians).

Table 1 gives an overview comparison between the generalized hierarchical search system and a typical state-of-the-art LVCSR system. The principal features of the generalized hierarchical search are described below.

Unlimited Context: An important technique in speech recognition is to model coarticulation effects through the use of context-dependent phones. In practice, the deeper the context used, the better the performance achieved. In 2001 Hub5 evaluations, for instance, one system achieved a 5% relative reduction in WER by using quinphones (20.2%) in place of their usual triphone system (21.3%) [3]. However, most systems allow only triphones or they require that special software be used to provide longer context support. The generalized speech system allows any length context at each level in the hierarchy.

Unlimited Hierarchical Levels: While most systems provide for a limited depth of knowledge, the generalized hierarchical search allows any number of decompositions. Each level is specified as a graph structure and the levels are combined together into one master graph structure as shown in Figure 1. Because the system is specified as a set of levels, each level has individual control parameters including the pruning approaches, context dependency and options for compression of the graphs at that level.

Dynamically Switched Language Models: A significant feature of the ISIP decoder is the ability to decode simultaneously using multiple language models (LMs). It can switch from one LM to another dynamically during runtime. This is ideal for applications involving a broad language model like an N-gram that covers the general interaction with the user, but requires more specific sub-models to decode certain parts of the user speech input. The sub-models are typically generated by defining grammar structures of the language used in the specific recognition task. These grammar structures efficiently incorporate the higher level knowledge to further reduce the size of the search space. Thus, the search space is shared by hypotheses that not only follow the N-gram LM, but also by those that are decoded in parts by a number of different finite-state context-free grammars.

Generalized Specification: The designation of search levels in the hierarchy are specified through a parameter file. This is a departure from typical LVCSR systems which require extensive code changes to add a new

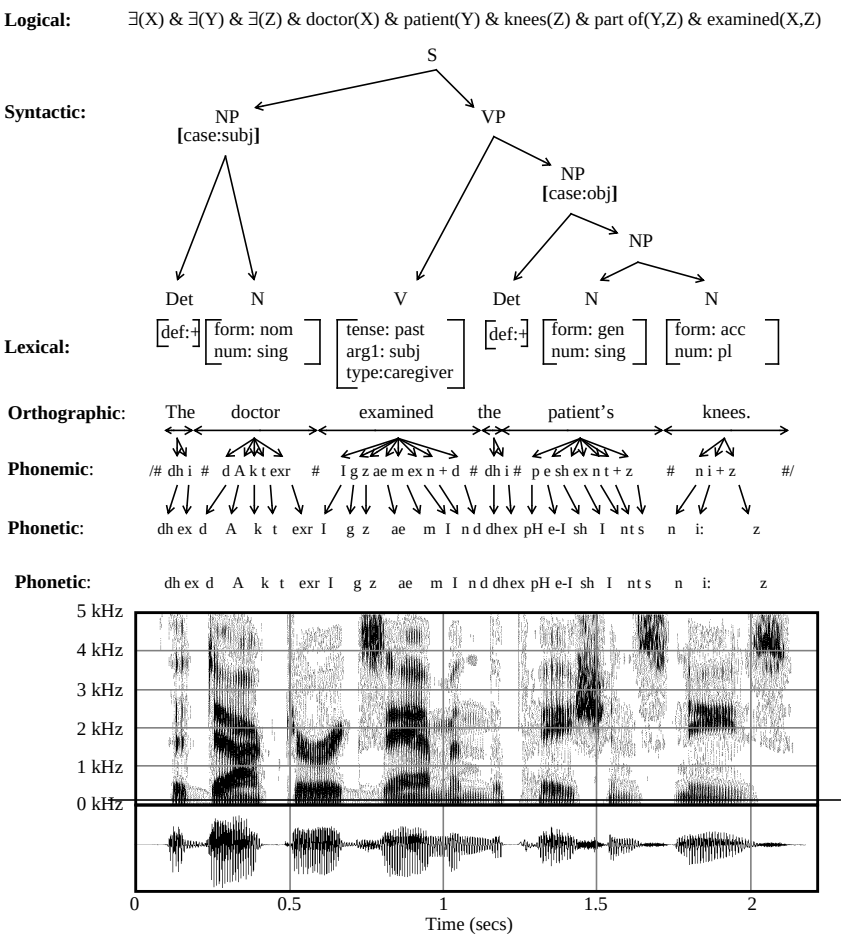

Figure 2: Language is composed of many hierarchical constraints which interact with one another. A generalized system is necessary to allow the incorporation of these constraints.

technique to the system. In this respect, the generalized approach provides an extremely powerful toolkit for new research.

Pruning and Search-space Compression: Search typically consists of finding the most likely path through the sequence of words. By employing dynamic programming under the linguistic constraints (as specified by the levels in the hierarchy) we can reduce the computational overhead in generating an immense number of hypotheses. However, it is still critical that we use pruning techniques to discard low-likelihood paths. The generalized system provides independent pruning specifications at each level in the hierarchy. These include:

1. **Viterbi Pruning:** Using the principle of dynamic programming, only the best path coming to the same point in the search space is preserved
2. **Beam Pruning:** Paths with scores that fall within some constant margin from the best score at each time frame and at each level are preserved.
3. **Active Instance Pruning:** The number of unique path instances can be specified. This effectively limits the total number of scoring slots necessary in the search process.
4. **Lexical Trees:** Two or more levels can be collapsed to reduce the total size of the search space [4]. Most LVCSR systems do this at the phoneme level to give an efficient form of the pronunciation lexicon.

3. EXPERIMENTS AND RESULTS

We have conducted several experiments on different corpora to demonstrate that the system [5, 6] is comparable to state-of-the-art systems. Table 2 shows small vocabulary experiments on the TIDIGITS corpus [7] using coded speech data. Two systems were tested: a 16 mixture HMM-based recognizer that uses whole word

Data	Word Error Rate	
	Word Models	Xwrd CD Models
STUDIO	0.4%	0.6%
MELP	0.7%	0.8%

Table 2: WER for TIDIGITS studio and coded data using word models and cross-word triphone context dependent models.

Task	Acoustic Model	Language Model	WER
RM	XWRD	Bigram	3.4%
WSJ	XWRD	Bigram	8.3%
Hub5e01	WINT XWRD	Bigram Trigram	35.6%

Table 3: Performances on the Resource management, Wall Street Journal and SWITCHBOAD corpora.

acoustic models, and a similar system that uses cross-word triphones for its acoustic models. The performance of these systems on the studio data was 0.4% and 0.6% WER respectively. Using coded data, we experience a slight degradation in performance.

We next developed a medium vocabulary system using the DARPA Resource Management corpus [8]. This system has a 1000 word vocabulary and a bigram language model with a perplexity of 60. Acoustic models were cross-word triphones with 6 Gaussian mixture components per state. A WER of 3.4% was achieved at 9.7 real-time rate using a 600 MHz Pentium processor. We were also able to tune the system to run at near real-time with a slight increase in WER to 5.0%.

Finally, we have created a large vocabulary system using the Wall Street Journal corpus [9]. This system is based on a state-tied cross-word triphone acoustic model with 16 Gaussian mixtures per state. Evaluation of the Eval'92 data set using a bigram language model provided by Lincoln Labs, gives a WER of 8.3% which is comparable to state-of-art systems.

4. CONCLUSIONS

LVCSR systems have advanced significantly over the last few years due to increase computational power and development of very efficient search algorithms. However, for most systems the integration of an additional knowledge source into the search is difficult if not impossible. We have introduced a system designed around a generalized hierarchical search that eliminates these significant drawbacks. We have described the flexibility of the search which makes it a powerful testbed for researchers to implement and test novel concepts.

REFERENCES

[1] N. Deshmukh, A. Ganapathiraju and J. Picone, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition," IEEE Signal Processing Magazine, vol. 16, no. 5, pp. 84-107, September 1999.

[2] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, Cambridge, Massachusetts, London, England, 1998.

[3] P. Woodland, T. Hain, G. Evermann and D. Povey, "CU-HTK March 2001 Hub5 System," 2001 Large Vocabulary Conversational Speech Recognition Workshop, Baltimore, Maryland, May 2001.

[4] H. Murveit, P. Monaco, V. Digalakis and J. Butzberger, "Techniques to Achieve an Accurate Real-Time Large-Vocabulary Speech Recognition System," Proceedings of the ARPA Human Language Technology Workshop, pp. 368-373, Austin, Texas, USA, March 1995.

[5] "ISIP AUTOMATIC SPEECH RECOGNITION SYSTEM," *http://www.isip.msstate.edu/projects/speech/software/asr/download/asr/index.html*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 2001.

[6] J. Picone, "Tutorials", http://www.isip.msstate.edu/projects/speech/software/tutorials/, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, November 2001..

[7] R. G. Leonard, "A Database for Speaker-Independent Digit Recognition," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 3, pp. 42.11, Dallas, Texas, USA, 1984.

[8] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-Word Resource Management Database for continuous speech recognition," IEEE International Conference on Acoustics, Speech, and signal Processing, pp. 651-654, 1988.

[9] D. Paul and J. Baker, The Design of Wall Street Journal-based CSR Corpus, Proceedings of the International Conference on Spoken Language Systems (ICSLP), pp. 899-902, Banff, Alberta, Canada, October 1992.