LOWERRE, BRUCE T.
THE HARPY SPEECH RECOGNITION SYSTEM

30

The HARPY Speech Recognition System

Thesis Summary

Bruce T. Lowerre

April, 1976

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Submitted to Carnegie-Mellon University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported in part by the Advanced Research Projects Agency of the Office of the Secretary of Defense (Contract F44620-73-C-0074) and monitored by the Air Force Office of Scientific Research. .

Abstract

The Harpy connected speech recognition system is the result of an attempt to understand the relative importance of various design choices of two earlier speech recognition systems developed at Carnegie-Mellon University: The Hearsay-I system and the Dragon system. Knowledge is represented in the Hearsay-I system as procedures and in the Dragon system as a Markov network with a-priori transition probabilities between states. Hearsay-I uses a best-first search strategy of the syntactic paths while Dragon searches all the possible syntactic (and acoustic) paths through the network in parallel to determine the globally optimal path. Hearsay-I uses segmentation and labeling to reduce the effective utterance length while Dragon is a segmentation-free system. Systematic performance analysis of various design choices of these two systems resulted in the HARPY system, in which knowledge is represented as a finite state transition network but without the a-priori transition probabilities. Harpy searches only a few "best" syntactic (and acoustic) paths in parallel to determine the optimal path, and uses segmentation to effectively reduce the utterance length, thereby reducing the number of state probability updates that must be done.

Several new heuristics have been added to the HARPY system to improve its performance and speed: detection of common sub-nets and collapsing them to reduce overall network size and complexity, eliminating the need for doing an acoustic match for all phonemic types at every time sample, and semi-automatic techniques for learning the lexical representations (that are needed for a steady-state system of this

type) and the phonemic templates from training data, thus automatically accounting for the commonly occurring intra-word coarticulation and juncture phenomena. Inter-word phenomena are handled by the use of juncture rules which are applied at network generation time, thereby eliminating the need for repetitive and time consuming application of phonological rules during the recognition phase. State transition probabilities are calculated dynamically during the recognition phase from speech dependent knowledge rather than a-priori from statistical measurements.

The system was trained on four similar sets of data that were recorded by four different speakers. Four sets of acoustic-phonetic templates were produced, one for each speaker, but only a single phonetic dictionary is used for all four speakers. The syntactic grammar used for testing the system is a desk calculator task (sometimes called "voice programming") that contains 37 words and has an average branching factor of about 11. The system was also tested on another grammar (over the same 37 words) that provided no syntactic support.[1] The Harpy system achieves 88.6% utterance accuracy and runs in 13.1 times real-time (on a PDP KA-10), with syntactic grammar constraints, on a set of training data that was recorded separately from but during the same session as its training data by the same four speakers. The system achieves 62.0% utterance accuracy in 18.0 times real-time on this same test data without syntactic constraint. On another set of test data from the same four speakers that was recorded five months after the training data, the system achieves 70.9% utterance accuracy in 13.2 times real-time with syntax and 36.7% utterance accuracy in 18.0 times real-time without syntax.

[1] I.e., a grammar specification is used that allowed any word to follow any other word. The branching factor in this case is 37.

Motivation

The original title of this thesis was "A Comparative Performance Analysis of Speech Understanding Systems". It was intended to be a study of the various types and levels of knowledge that are used in speech understanding systems (e.g., parametric schemes, acoustic-phonetic evaluation functions, and search strategies) to determine their relative strengths and weaknesses. One of the objectives of the original thesis was, after the analysis study, to combine the "good" features of the systems that were studied to form a system that would have performance characteristics superior to the others. However, the work on this "hybrid" system proved to be more interesting than the original analysis study.

The Harpy system evolved from a systematic study of several design choices of two speech systems previously developed at Carnegie-Mellon University: The Hearsay-I system (see Erman (1974)) and the Dragon system (see Baker (1975)). These two systems are of interest because of their contrasting features: model, structure, performance, and representation of knowledge. The following table shows some of these features:

FEATURE	HEARSAY-I	DRAGON
Model	cooperating parallel processes	probabilistic function of a Markov process
knowledge representation	procedures	Markov network
units of speech signal representation	segmentation	10 milli-second sample

search strategy	best-first	all paths in parallel
search time	8-50 times real-time	45-200 times real-time
search time variation	extremely large	near zero
performance	does not always find optimum recognition path. Occasionally results in no recognition	always finds optimum recognition path
syntax	uses anti-productions of a BNF grammar to generate the syntactic paths to search	syntax is an integral structure of the Markov network
phonetics	uses a phonetic dictionary and many "hard-wired" heuristics to do the phonetic parsing	phonetic pronunciations are an integral structure of the Markov network. contains no other phonetic knowledge
acoustics	uses user dependent phonetic templates and many "hard-wired" heuristics for evaluating the acoustic signal	uses user dependent phonetic templates and a simple probabilistic function for evaluating the acoustic signal

These design choices were carefully studied to determine their relative strengths and weaknesses.

Goals of the Harpy System

The goals of the Harpy system are to combine the "best" features of the above two systems with additional heuristics to form a high speed and high performance system. The features of the Harpy system are:

FEATURE	HARPY
Model	Dynamic programming system
knowledge representation	state transition network
units of speech signal representation	segmentation
search strategy	a "few best" paths in parallel
search time	about 10-20 times real-time
search time variation	near zero
performance	nearly always (i.e., more than 99% of the utterances) finds the optimum recognition path
syntax	syntax is an integral structure of the state transition network.
phonetics	Phonetic pronunciations are an integral structure of the start transition network. Word juncture phenomena are also an integral structure of the network. Contains no other phonetic knowledge.
acoustics	Uses user dependent phonetic templates and a simple probabilistic function for evaluating the acoustic signal. Also contains phonetic duration heuristics for calculating state transition probabilities.

Several heuristics and techniques were developed to reduce the network size and complexity. Also, new techniques were developed for the semi-automatic generation (and tuning) of the lexicon and templates that are needed for this type of "steady-state" system. This tuning has resulted in a vast improvement of the accuracy of the system.

Features of the Hearsay-I system

The Hearsay-I system is the result of an attempt to represent the diverse sources of knowledge needed for speech recognition as procedures. The system is highly modularized into separate knowledge source procedures. This modularity allows each knowledge procedure to contain highly specialized heuristics about its own knowledge. The Hearsay-I system is the product of about 10 man-years of effort from about five principal contributors. The author is responsible for adding a large amount of the acoustic-phonetic knowledge to this system. Historically, this was the first system to be demonstrated live.

Model of the Hearsay-I system

The model of the Hearsay-I system is one of parallel cooperating processes. The system contains three of these parallel processes (modules): semantics, syntax, and acoustics. However, for the purposes of comparison with Harpy, only syntax and acoustics are considered. The search strategy used in the Hearsay-I system is a best-first search of the legal syntactic paths (at the word level) with backtracking. The search is driven by syntactic anti-productions which hypothesize words. The acoustics module verifies these words by either rating them or rejecting them. The search of the syntactic paths is done in a best-first strategy. See Reddy et al.(1973) and Erman (1974) for a detailed example and description of this system.

Discussion of the Hearsay-I system

The features of the Hearsay-I system that were considered in designing the Harpy system are the following:

1) Segmentation of the acoustic signal can effectively reduce the amount of speech data searched.

2) A fast recognition can be achieved by searching a small number of paths (e.g., one in the case of the best-first search).

3) Backtracking while searching can be costly, especially in a large search space.

4) Duplication of effort should be avoided (i.e., knowledge gained from past searching should be used to guide future searching).

5) Heuristic speech knowledge is a useful guide to mapping and rating words.

Features of the Dragon system

The Dragon system was essentially a one-man effort and represents about two years of work. The system is interesting because of its simplicity of design, its mathematical tractability, and its high accuracy performance. The program code contains almost no speech dependent heuristics. The author is responsible for tuning this system to the point where it is the first system developed at Carnegie-Mellon University to achieve 100% accuracy on a set of speech data using only syntactic and acoustic-phonetic knowledge.

Model of the Dragon system

The model of the Dragon system (See Baker (1975)) is one of a probabilistic function of a Markov process. The system achieves recognition by updating state probabilities of a Markov network on every 10 milli-second speech sample. The network contains all the syntactic and phonetic knowledge used in the system. Inter-state connections are indicated by a-priori transition probabilities. The network also contains intra-state a-priori transition probabilities.

Discussion of the Dragon system

The Dragon systems demonstrates that it is possible to build a connected speech recognition system that achieves better than 85% word accuracy using only syntactic and acoustic knowledge. More importantly, the system also shows that there is a search algorithm (that does no backtracking) which guarantees a recognition and in a deterministic amount of time.

The most significant feature of the Dragon system, as compared to most other current speech recognition systems, is its almost total lack of speech-dependent heuristic knowledge. Dragon treats speech recognition as a mathematical computation problem rather than as an artificial intelligence problem.

Features of the Harpy system

The Harpy system is an attempt to combine the best features of the Hearsay-I system and the Dragon system. Harpy contains a mathematically tractable model, as in the Dragon system, plus speech-dependent heuristics, as in the Hearsay-I system, for better performance and speed than either of the other two systems. Harpy also incorporates many new techniques (and heuristics) for increasing its speed and accuracy. Also, an entirely new semi-automatic system was designed to generate the lexicon and templates needed for Harpy. The following is a comparison of the two previous systems and Harpy:

	HEARSAY-I	DRAGON	HARPY
Knowledge Representation	Procedural embedding	Markov networks	Transition networks (No a-priori transition probabilities)
Search strategy	best first with backtracking	all paths in parallel with no backtracking	"best few" in parallel with no backtracking
Segmentation	yes	no	yes
Phonetic classification	procedural	multiple templates	unique templates
Word Juncture Knowledge	procedural	none	integral part of network

Model of the Harpy system

The model of the Harpy system is one of a dynamic programming system with the use of heuristics to reduce the search space, thereby increasing its speed. The system uses a network that represents not only all legal syntactic paths, but also all pronunciations of these legal paths. The combined knowledge of syntax and lexicon is contained within all the legal "syntactic-phonetic" paths of the network. The network contains inter-state connections, but no a-priori transition probabilities.

Segmentation

Reduction of the number of time samples used to update the state probabilities is accomplished by segmentation of the input data into larger sized units than the 10 milli-second time samples. The only critical problem involved with doing segmentation for this type of recognition system is that there must be <u>no</u> missing segments. The system will map one or more time samples to one network state but it will never map more than one state to a single time sample. If the global path for an utterance of U 10 milli-second time samples has P states from initial state to final state (P≤U), then a segmentation scheme may produce anywhere between P and U number of segments. If it produces fewer than P segments, the global path cannot be parsed and therefore will not be recognized. The fewer segments that are produced (but greater than P, of course), the faster will be the recognition speed. One measure of efficiency of a segmenter is how few segments it produces. However, any errors that it makes must be on the side of too many rather than too few segments.

The use of segments requires a small change in the recognition algorithm. Since the segments produced are not of uniform length, the state acoustic-phonemic match probabilities and the intra-state transition probabilities must be weighted by the length of the current segment during the recognition process.

The segmentation algorithm used in the Harpy system combines contiguous 10 milli-second speech samples which match to each other within a heuristic threshold to form a segment. The parameters representing each segment are an average of the parameters of the samples comprising the segment.

The ratio of 10 milli-second samples to segments that the segmenter produces is dependent on the acoustic signal. On the average, it is about 3.5. The overhead for the segmenter takes 1.3 times real-time (i.e., processing time is 13 milli-seconds per 10 milli-second sample). The segmentation scheme gives the Harpy system better than a three fold speed up.

Search Space Reduction

The speed up philosophy of the Harpy system is to use heuristics to reduce the calculations and search space as much as possible during the recognition process. To accomplish this, Harpy uses a dynamic programming scheme to search a network

Performance of the Harpy system

After the system was tuned on its training data, it was run on its test data. The test data consist of two large sets each of which contains four smaller sets, one from each of four speakers. All three sets spoken by each of the four speakers (one training set and the two test sets) contain the same 20 sentences and 110 words. The first test set was recorded during the same session as the training set; the second test set was recorded five months after the training set. No analysis has been done on any of the errors made by the system on the test data, nor has any attempt been made to correct these errors. Error corrections were done only on the training set. Each test set was run in two modes: with syntactic support and without syntactic support. The network for the latter mode was created by using a BNF grammar that allowed any word to follow any other word. The effective syntactic branching factor (the average number of words that can follow any other word in a sentence) for the runs with syntax is about 11, and without syntax it is 37. See Goodman (1976) for a complete treatise on the complexities of these two grammars. The following are the overall results of the Harpy system on the two sets of test data:

TEST DATA	%WORDS	%UTTS	TIME
1) WITH SYNTAX	97.5	88.6	13.1
1) NO SYNTAX	90.8	55.0	17.8
2) WITH SYNTAX	93.1	70.9	13.2
2) NO SYNTAX	83.5	36,7	18.0

"1)" and "2)" indicate the immediate and five month test sets respectively. %WORDS is representing all legal syntactic paths and all pronunciations of these legal paths but searches only a few "best" paths in parallel. "Best" is defined by a heuristic threshold on the probabilities of the states being searched. This allows the system to search a variable number of paths; many paths where they become confusable and few where the correct path is "obvious". This reduction in search space results in a speed up of between 10 to 25 over searching all paths in parallel (as in the Dragon system).

Data Dependent Transition Probabilities

The networks used by the Dragon system contain a-priori constant transition probabilities that were calculated from statistical evidence. These constant probabilities can cause erroneous transitions since some transitions may be more likely than others. Harpy uses transition probabilities that are calculated dynamically during the recognition process from data dependent knowledge (specifically, intrinsic phone durations). This greatly reduces the gross errors produced from using constant transition probabilities.

Network Size Reduction

Several heuristics were developed to reduce the state size and complexity of the networks. These heuristics are removal of null states, removal of redundant states, and subsumption of common states. These heuristics can reduce the state size of the networks by up to 50%.

% words correctly recognized, %UTTS is the % utterances correctly identified (i.e., all words within the utterance are correctly recognized), and TIME is the number of times real-time needed for recognition. TIME includes 8.4 times real-time for the generation of the autocorrelation coefficients and the linear predictor coefficients.

REFERENCES

Baker, J.K. (1975), "The DRAGON System -- An Overview", IEEE Trans. ASSP-23, 24-29.

Baker, J.K. (1975), "Stochastic Modeling as a Means of Automatic Speech Recognition", (Ph.D. Thesis, Carnegie-Mellon Univ.), Tech Rep., Comp.Sci. Dept, Carnegie-Mellon University.

Baker, J.K. and Bahl, L. (1975), "Some Experiments in Automatic Recognition of Continuous Speech", Proceedings Eleventh Annual IEEE Computer Society Conference, 326-329.

Erman, L.D. (1974), "An Environment and System for Machine Understanding of Connected Speech", (Ph.D. Thesis, Stanford Univ.), Tech. Rep., Comp. Sci. Dept., Carnegie-Mellon University.

Erman, L.D. (ed.) (1974), Contributed Papers of IEEE Symposium on Speech Recognition, Carnegie-Mellon Univ., (IEEE cat. NO. 74CH0878-9 AE).

Erman, L.D., Lowerre, B.T., and Reddy, D.R. (1973), "Representation and Use of Acoustic-Phonetic Knowledge in the HEARSAY System", 86th Mtg Program, Acous. Soc. Am., Los Angeles, 49 (abstract).

Forgie, J.W., Hall, D.E. and Wiesen, R.W. (1974), "An Overview of the Lincoln Laboratory Speech Recognition System", J. Acoustic Society of Amer., 56, S27 (A).

Goldberg, H.G. (1975), "Segmentation and Labeling of Speech: A Comparative Performance Evaluation", (Ph.D. Thesis, Carnegie-Mellon Univ.), Tech. Rep., Comp. Sci. Dept., Carnegie-Mellon University.

Goldberg, H.G., Reddy, D.R., and Suslick, R. (1974), "Parameter-Independent Segmentation and Labeling of Speech", in Erman (ed.).

Goodman, G. (1976), "Language Design for Man-Machine Communication", (Ph.D. Thesis), Comp. Sci. Dept., Carnegie-Mellon University, (in preparation).

Hopcroft, J. E. and Ullman, J. D. (1969), Formal Languages and their Relation to Automata, Addison-Wesley Publishing Co., Mass.

Itakura, F. (1975), "Minimum Prediction Residule Principle Applied to Speech Recognition", IEEE Trans. ASSP-23, 67-72.

Karp, R.M., Miller, R.E., and Rosenberg, A.L., "Rapid Identification of Repeated Patterns in Strings, Trees and Arrays", IBM T. J. Watson Research Center, Yorktown Heights, N.Y.

Lowerre, B.T. (1974), "A Camparison of Two Speech Understanding Systems", 88th Mtg Program, Acous. Soc. Am., St. Louis, 27 (abstract).

Markel, J. (1972), "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. AU-20.

Markel, J.D. and Gray, A.H. Jr. (1973), "SCRL Linear Prediction Analysis/Synthesis Programs", SCRL, Inc., Santa Barbara, Calif.

Neely, R.B. (1973), "On the Use of Syntax and Semantics in a Speech Understanding System", (Ph.D. Thesis, Stanford Univ.), Tech. Rep., Comp. Sci. Dept., Carnegie-Mellon University.

Newell, A. (1975), "A Tutorial on Speech Understanding Systems", in Reddy (ed.).

Newell, A., Barnett, J., Forgie, J., Green, C., Klatt, D., Licklider, J.C.R., Munson, J., Reddy, D.R., and Woods, W. (1971), "Speech Understanding Systems: Final Report of a Study Group", Reprinted by North-Holland/American Elsevier, Amsterdam.

Reddy, D.R. (ed.) (1975), Speech Recognition, Invited Papers of the IEEE Symposium, Academic Press, N.Y.

Reddy, D.R. (1976), "Speech Recognition by Machine: A Review", to be published, Proc. IEEE, April.

Reddy, D.R., Erman, L.D., and Neely, R.B. (1973), "A Model and a System for Machine Recognition of Speech", IEEE Trans. AU-21, 229-238.

Reddy, D.R., Erman, L.D., Fennell, R.D., and Neely, R.B. (1973), "The HEARSAY-I Speech Understanding System: An Example of the Recognition Process", Proc. 3rd Inter. Joint Conf. on Artificial Intelligence, Stanford, Ca., 185-193, to appear in IEEE Trans. Computer, April, 1976.

VanLehn, K.A. (1973), SAIL Users Manual, Report STAN-CS-73-373, Computer Science Department, Stanford Univ., Calif.

Vicens, P. (1969), "Aspects of Speech Recognition by Computer", (Ph.D. Thesis, Stanford Univ.), Rept. CS-127, Comp. Sci. Dept., Stanford Univ.