THE HEARSAY SPEECH UNDERSTANDING SYSTEM: An Example of the Recognition Process

D.R. Reddy, L.D. Erman, R.O. Fennell, and R.B. Neely*

Computer Science Department**
Carnegie -Mellon University
Pittsburgh, Pa. 15213

ABSTRACT

This paper describes the structure and operation of the Hearsay speech understanding system by the use of a specific example illustrating the various stages of recognition. The system consists of a set of cooperating independent processes, each representing a source of Knowledge. The knowledge is used either to predict what may appear in a given context or to verify hypotheses resulting from a prediction. The structure of the system is illustrated by considering its Operation in a particular task situation: Voice-Chess. The representation and use of various sources of knowledge are outlined. Preliminary results of the reduction in search resulting from the use of various sources of knowledge are given.

Keywords: speech recognition, understanding, hypothesize-and-test

The factors influencing the structure and operation of a speech understanding system are many and complex. The report of Newell et al. (1971) discusses these issues in detail. Our own goals and efforts in this area have been described in several earlier papers (Reddy et al., 1972). The goals for our present effort were outlined in Reddy, Erman, and Neely (1970). The initial structural description of the Hearsay system was given in Reddy (1971). The model and the system that evolved after several design iterations were described in Reddy, Erman, and Neely (1972a).* The main additions to the initial proposed system were *in* the specification of the interactions among various sources of knowledge. In this paper, we describe the structure and operation of the Hearsay system from a different point of view, i.e., by considering a specific example to illustrate the various stages of the recognition process.

Machine perception of speech differs from many other problems in artificial intelligence in that it is characterized by high data rates, large amounts of data, and the availability of many sources of knowledge. Thus, the techniques that must be employed differ from other problem-solving systems in which weaker and weaker methods are used to solve a problem using less and less Information about the actual task. In addition, there is a marked difference in the expectations for system performance. In tasks such as chess and theorem-proving, the human has sufficient trouble himself so as to make reasonably crude programs of interest. But humans perform effortlessly (and with only modest error) in speech or visual perception tasks, and they demand comparable performance from a machine. Thus, it is important that the structure and organization of a system be such that it is not a dead-end effort, i.e., it should be capable of approaching human performance without major reformulation of the problem solution. The Hearsay system effort represents an attempt to produce one such system. The main distinguishing characteristic of this system is that diverse sources of knowledge can be represented as cooperating independent parallel processes which help in the decoding of the utterances using the hypothesize-and-test paradigm.

The system is designed for the recognition of connected speech, from several speakers, with graceful error recovery, performing the recognition in close to real-time. The structure and implementation of the system are to a large extent dictated by these concerns. One feature that characterizes a speech understanding system is the existence of errors at every level of analysis. The errorful nature of processing Implies that every source of knowledge has to be invoked to resolve ambiguities and errors at every stage of the processing. One way to accomplish this is through the use of the hypothesize-and-test peradjgm, where each *source* of knowledge *can accept,* reject, *or* re-order the hypotheses produced by other sources of knowledge. For example, in the Voice-Chess task, if the word "captures" appears in a partially-recognized utterance, the

* The general framework that evolved for the model is different from some previously proposed models by Liba man et al. (1962) and Halle and Stevens (1962) which imply that perception takes place through the active mediation of motor centers. Our efforts tend to support "sensory" theories advanced by Fant (1964) and others. If one modifies the "synthesis" part of analysis-by-synthesis, then our model is most similar to that of Halle and Stevens.

• Present address: Xerox Palo Alto Research Center, Palo Alto, Ca. 94305.

** This research was supported in part by the Advanced Research Projects Agency of the Department of Defense under contract no. F44620-70-C-0107 and monitored by the Air Force Office of Scientific Research.

semantic source of knowledge can reject all the hypotheses that do not lead to a capture move.

The Hearsay system is not restricted to any particular recognition task. Given the syntax and the vocabulary of a language and the semantics of the task, it attempts recognition of utterances in that language. It is designed to serve as a research tool in which the contributions of various sources of knowledge towards recognition can be clearly evaluated. Since each source of knowledge is represented as an Independent process, it can be removed without crippling the system.

Figure 1 gives an overview of the Hearsay system. The EAR module accepts speech input, extracts parameters, and performs some preliminary segmentation, feature extraction and labeling, generating a "partial symbolic utterance description." ROVER (Recognition OVERlord) controls the recognition process and coordinates the hypothesis generation and verification (testing) phases of the various cooperating knowledge processes. The TASK provides the interface between the task being performed and the speech recognition and generation (SPEAK-EASY) parts of the system. SOL, the System Over Lord, provides the message communication facilities for the system.

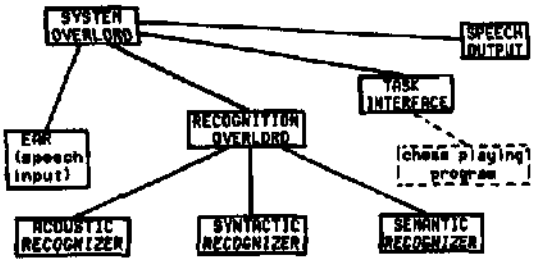


Figure 1: Structure of the Hearsay system.

AN EXAMPLE OF RECOGNITION

Here we will illustrate the operation of the Hearsay system by considering in detail the recognition process of an utterance within a specific task environment; Voice-Chess. The task is to recognize a spoken chess move in a given board position and respond with the counter-move.

Figure 2 gives the board position and a list *of* legal moves in that position at the time the move is spoken. The speaker, playing white, wishes to move his bishop on queen's bishop one to king knight five. This is one of 46 different legal moves. These moves have been ordered on the basis of their goodness in the given board position. This judgment was based on a task-dependent source of knowledge available to the program (Giltogly, 1972). Note that the move chosen by the speaker was only the fourth best move in that situation.

Having chosen the move, there are many possible ways of uttering the move. The syntax of the language permits many variations, usually of the form <piece> <action> <posltion>. The piece can have qualifiers to indicate the location. The action may be of the form: "to", "moves-to", "goes-to", "takes", "captures", and ao on. The position can be of the form: "king three", "king bishop four", or "queen's knight five", and so on. The actual move spoken in this context was "bishop moves-to king knight five". Note that "queen bishop on queen bishop one" can be specified as just "bishop" because there is no ambiguity In this case.

Figure 3 shows the speech waveform of the utterance with manual segmentation, showing the beginning and ending of each word and each phoneme within the word. (The menal

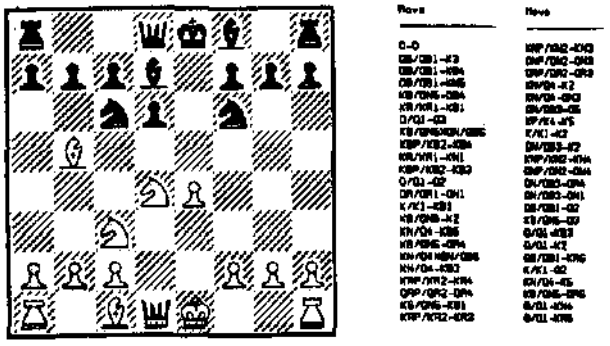


Figure 2: The chess board position and the ordered list of legal moves for White.

segmentation and labeling indicated in this and succeeding figures is for our benefit only — it is not available to the system while it is attempting recognition.) The utterance was about 2 seconds in duration and the waveform is displayed on ten consecutive rows, each row containing 200 milliseconds of the utterance. The first line of text under each row contains the word being articulated. The word label is repeated for the duration of the word. Thus, the word "bishop" was articulated for 400 milliseconds and occupies the first two rows of the waveform. The Sec*ond* line of text under each row contains the intended phoneme being articulated. The phoneme (represented in 1PA notation) Is repeated for the duration of the phoneme.

Several interesting problems of speech recognition arise in the context of recognition of this utterance. The end of Row 2 of Figure 3 shows the juncture between "bishop" and "moves". Note that the ending /p/ in "bishop" and the beginning nasal /m/ in "moves" are homorganic, i.e., they both have the same articulatory position. This results in the absence of the release and the aspiration that normally characterizes the sound /p/. Row 6 of Figure 3 illustrates a word boundary problem. The ending nasal of "king" and the beginning nasal of "knight" tend to be articulated from the same tongue position even though in isolation they would have been articulated from two different positions. This results in a single segment representing two different phonemes in two adjacent words. Further, It is impossible to specify the exact location of the word boundary. In the manual segmentation, the boundary was placed at an arbitrary position. Another type of juncture problem appears on Row 8 of Figure 3 at the boundary of "knight five". The release and aspiration of the phoneme /t/ are assimilated into the /f/ of "five".

Feature. Extraction and Sementation

The speech input from the microphone is passed through five band-pass filters (spanning the range 200-6400 Hz) and through an unfiltered band. Within each band the maximum intensity is measured for every 10 milliseconds (the zero crossings are also measured in each of the bands but they do not play an important role in the recognition process at present). This results in a vector of 6 amplitude parameters every 10 milliseconds. These parameters are smoothed and log-transformed. Figure 4 shows a plot of these parameters as a function of time for part of the utterance of Figure 3. The top line shows the utterance spoken. The second line of text indicates where the word boundaries were marked during the manual segmentation process (this will permit manual verification of the accuracy of the mechine recognition process in the later stages).

This vector of parameters (labeled 1, 2, 3, 4, 5, and U In Figure 4) is, for each centisecond, compared with a standard set of parameter vectors to obtain a minimum distance classification

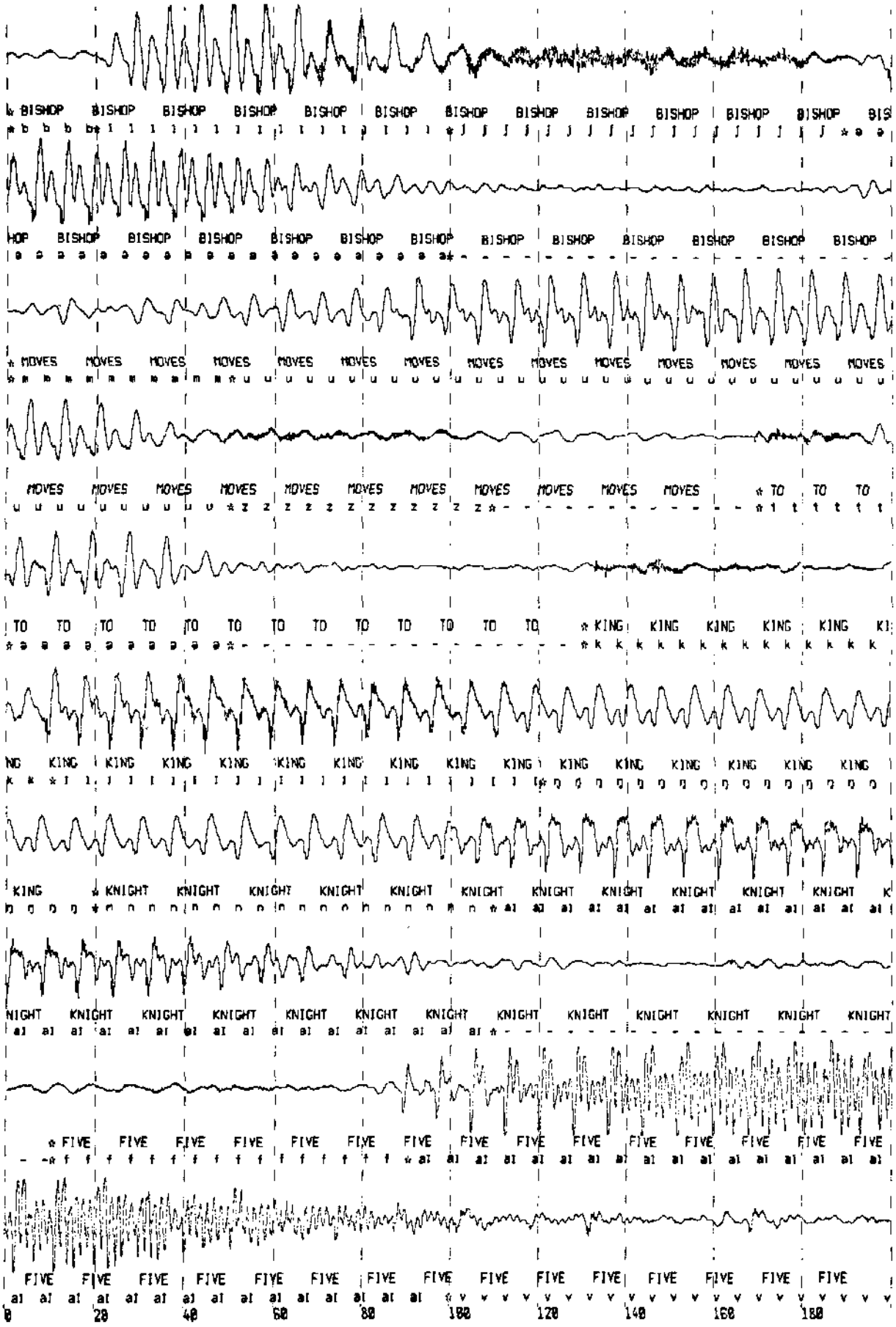


Figure 3: Waveform of the utterance with the "actual" word and phoneme boundaries.

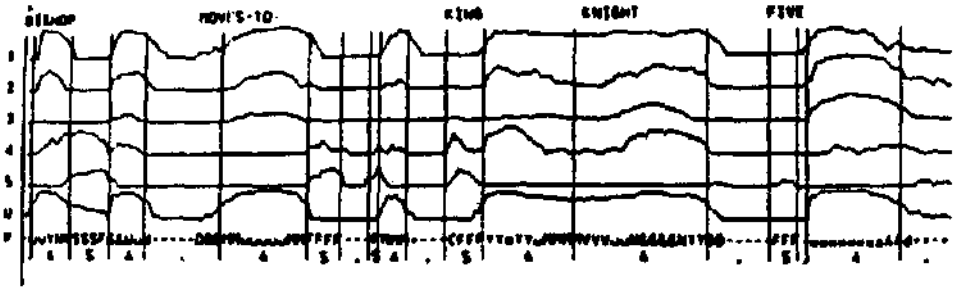


Figure 4. Parametric representation of the utterance showing the results of feature extraction and segmentation

using a modified nearest-neighbor classification technique. The purpose of this operation is to assign a (single character) label to each centisecond of speech using a compact pseudo-phonetic notation representing the actual local characteristics of the speech signal. The line of text labeled P in Figure 4 gives the classification for every 10-millisecond unit.

The classification of labels for each centisecond obtained by this match procedure (row P in Figure 4) is then used to specify a list of features, such as voicing and frication, which are then used in the segmentation of the utterance, shown in Figure 4. The boundaries of segments are indicated by vertical lines through the parameters, and the letter at the center of each segment (following the row P in Figure 4) indicates the type of segment that is present. The "A" indicates a sonorant segment, i.e., all the voiced unfricated segments; the "S" Indicates a fricated segment, and the period {"."} indicates a silence segment. The first use of an acoustic-phonetic source of knowledge can be seen *in* the handling of the "King knight" *word boundary* problem mentioned earlier. A long sonorant segment is subdivided into two segments to indicate the presence of two different syllables. The syllable juncture is determined in this case by the presence of a significant local minimum in an overall intensity plot (line labeled U on Figure 4).

The Recognition Process

The Hearsay system, at present, has three cooperating independent processes which help in the decoding of the utterances. These represent acoustic, syntactic, and semantic sources of Knowledge:

1. The acoustic-phonetic domain, which we refer to as just acoustics, deals with the sounds of the language and how they relate to the speech signal produced by the speaker. This domain of knowledge has traditionally been the only one used in most previous attempts at speech recognition.
2. The syntax domain deals with the ordering of words in the utterance according to the grammar of the input language.
3. The semantic domain considers the meaning of the utterances of the language, in the context of the task.

The actual number and nature of these sources Of knowledge is somewhat arbitrary. What is important to notice is that there can be several cooperating independent processes.

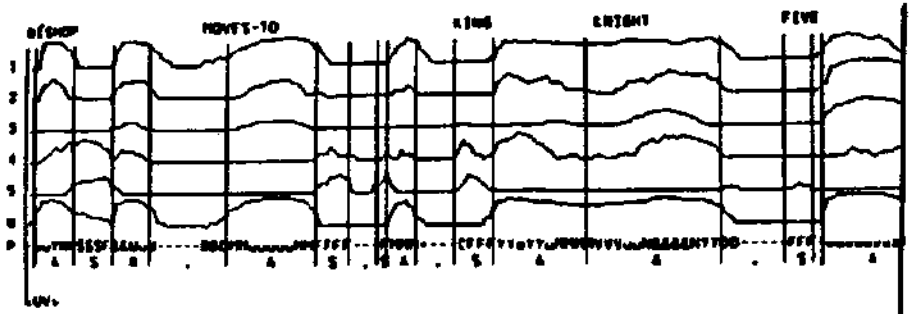
These processes cooperate by means of a hypothesize-end-test paradigm. This paradigm consists of one or more sources of knowledge looking at the unrecognized portion of the utterance and generating an ordered list of hypotheses. These hypotheses *may* then be verified by one or more of the sources of knowledge; the verification may accept, reject, or re-order the hypotheses. The same source of knowledge may be used in different ways both to generate hypotheses and to verify (or reject) hypotheses.

We will illustrate this recognition process by following through various stages of recognition for the utterance given in Figures 3 and 4. Figures 5 through 12 illustrate several of these stages of the recognition. In each figure, we have four kinds of information in addition to what was shown in Figure 4: the current sentence hypothesis (immediately below the P and segmentation rows), the processes acting on the current sentence hypothesis and their effect (e.g., SYN HYPOTHESIZED..., AGO REJECTED...), the acceptable option words with their ratings and word boundaries (e.g., I...I 500 Rook's), and the four best sentence hypotheses which result by adding the possible option words to the current best sentence hypothesis. When there are more than eight option words, only the best eight are shown. When there are more than four sentence hypotheses, only the best four are shown. The symbol <UV> within the current sentence hypothesis gives the location of the set of new words being hypothesized and verified. The "T...T" arrows indicate the possible beginning and ending for each option word.

Figure 5 shows the first cycle of the recognition process. At this point none *oi* the words in the sentence have been recognized and the processing begins left to right. The Syntax module chooses to hypothesize and generates 13 possible words, implying that the sentence can begin with "rook's", "rook", "queen's", etc. Of these, the Acoustics module absolutely rejects the word "bishop's" as being severely inconsistent with the acoustic-phonetic evidence. The Semantics module rejects "castle" and "castles" as being illegal in this board position. The remaining 10 words are rated by each of the sources of knowledge. The composite rating and the word beginning and ending markers for the eight best words are shown in Figure 5. The *words "rook",* "rook's", "queen's" *and* "queen" all get a rating Of 500. "Bishop", the correct word, gets a rating of 513. These words are then used to form the beginning sentence hypotheses, the top four of which are shown at the bottom of Figure 5.

Figure 6 shows the second cycle of the recognition process. The top sentence hypothesis is "bishop —". An attempt is being made to recognize the word following "bishop". Again Syntax generates the hypotheses. Given that "bishop" is the preceding word, the syntactic source of knowledge proposes only 7 options out of the possible 31 words in the lexicon — a reduction in search space by a factor of 4. Of these possible 7 words, Acoustics rejects "captures" and Semantics rejects none. The remaining six words are rated by each of the sources of knowledge and a composite rating along with word boundaries Is shown In Figure 6 for each of the acceptable words ("to" has a rating of 443, etc.). The correct word, "moves-to", happens to get the highest rating of 525. The new top sentence hypothesis is "bishop moves-to —", with a composite sentence rating of S47.

Figure 7 shows the third cycle of the recognition process.

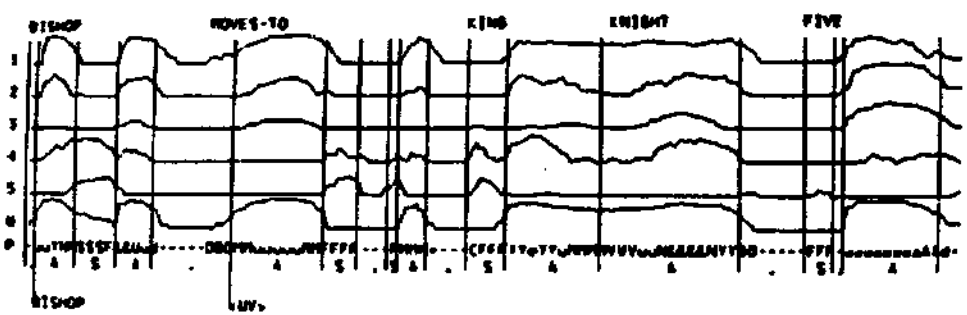



```
!.......! 500 ROOK'S
!.......! 500 ROOK
!.......! 500 QUEEN'S
!....! 500 QUEEN
!....! 425 PAWN
!....! 483 KING
! 513 BISHOP.....!
!.......! 483 KING'S

SYN HYPOTHESIZED    13 OW'S: ROOK'S,ROOK,QUEEN'S,QUEEN,PAWN,KNIGHT'S,KNIGHT,KING'S,KING,CASTLES,
ACO REJECTED         1 OW'S: BISHOP'S
SEM REJECTED         2 OW'S: CASTLES,CASTLE

THE TOP 4 SENTENCE HYPOTHESES ARE:
( 553) BISHOP <FILLER>
( 528) ROOK'S <FILLER>
( 526) ROOK <FILLER>
( 526) QUEEN'S <FILLER>
```

Figure 5: First stage of the recognition process.

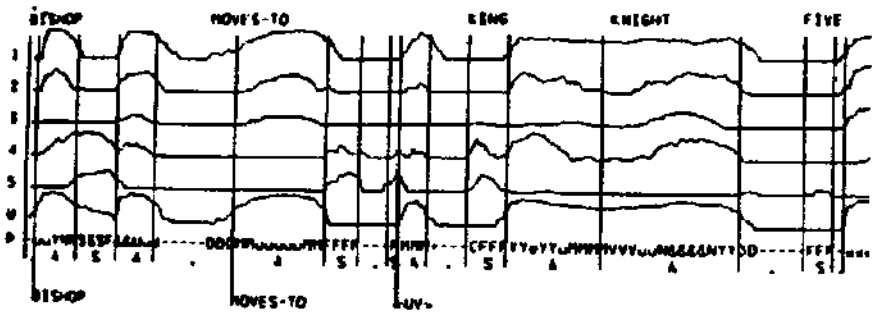



```
! 443 TO.....!
! 545 TAKES.....!
! 455 PAWN.....!
      ! 435 ON.!
      ! 525 MOVES-TO...!
      ! 468 GOES-TO...!
      !   0 CAPTURES !

SYN HYPOTHESIZED    7 OW'S: TO,TAKES,PAWN,ON,MOVES-TO,GOES-TO,CAPTURES
ACO REJECTED        1 OW'S: CAPTURES

THE TOP 4 SENTENCE HYPOTHESES ARE:
( 547) BISHOP MOVES-TO <FILLER>
( 553) BISHOP <FILLER>
( 528) BISHOP GOES-TO <FILLER>
( 523) BISHOP PAWN <FILLER>
```

Figure 6: Second stage of the recognition process.

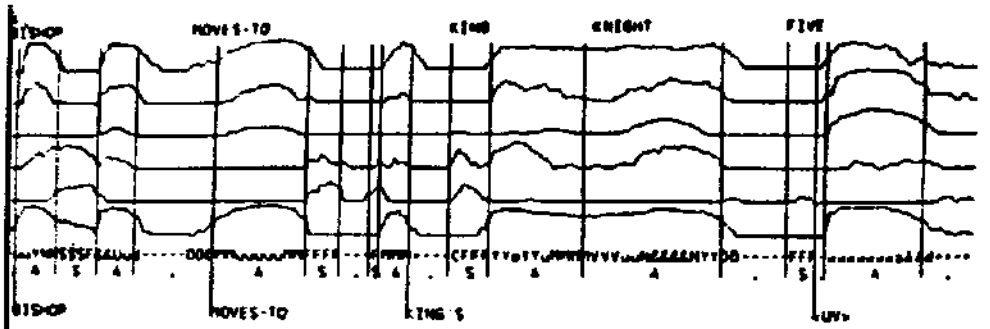



```
!.....! 458 ROOK
   ! 503 QUEEN'S.....!
   ! 503 QUEEN.....!
!.....! 348 KNIGHT
   ! 513 KING'S.....!
   ! 513 KING.....!
! 423 BISHOP.....!

SYN HYPOTHESIZED    7 OW'S: ROOK,QUEEN'S,QUEEN,KNIGHT,KING'S,KING,BISHOP

THE TOP 4 SENTENCE HYPOTHESES ARE:
( 552) BISHOP MOVES-TO KING'S <FILLER>
( 552) BISHOP MOVES-TO KING <FILLER>
( 547) BISHOP MOVES-TO <FILLER>
( 547) BISHOP MOVES-TO QUEEN'S <FILLER>
```

Figure 7: Third stage of the recognition process.




```
                                    ! 388 ROOK.....!
                                    ! 378 KNIGHT...!
                                    !   0 BISHOP !

SYN HYPOTHESIZED    3 OW'S: ROOK,KNIGHT,BISHOP
ACO REJECTED        1 OW'S: BISHOP

THE TOP 4 SENTENCE HYPOTHESES ARE:
( 552) BISHOP MOVES-TO KING'S <FILLER>
( 552) BISHOP MOVES-TO KING <FILLER>
( 547) BISHOP MOVES-TO <FILLER>
( 547) BISHOP MOVES-TO QUEEN'S <FILLER>
```

Figure 8: Fourth stage of the recognition process.

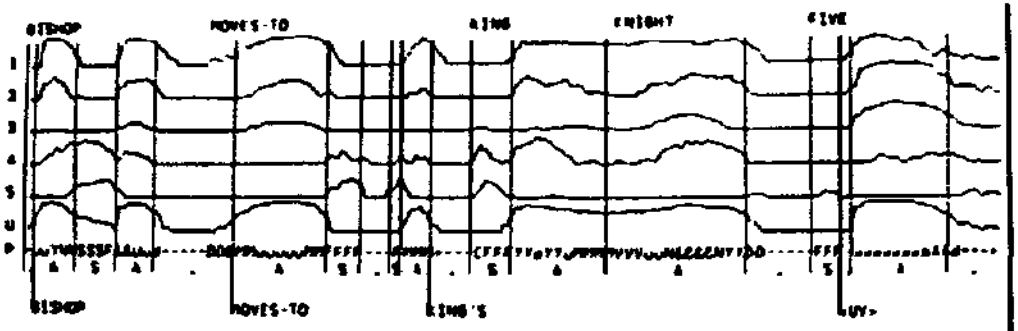


Figure 9: Fifth stage of the recognition process.

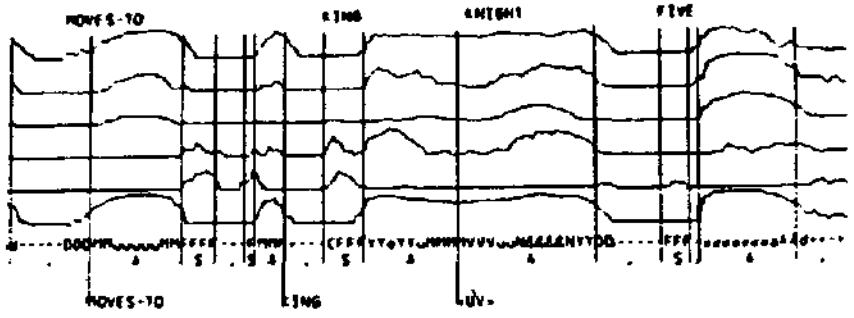


Figure 10: Sixth stage of the recognition process.

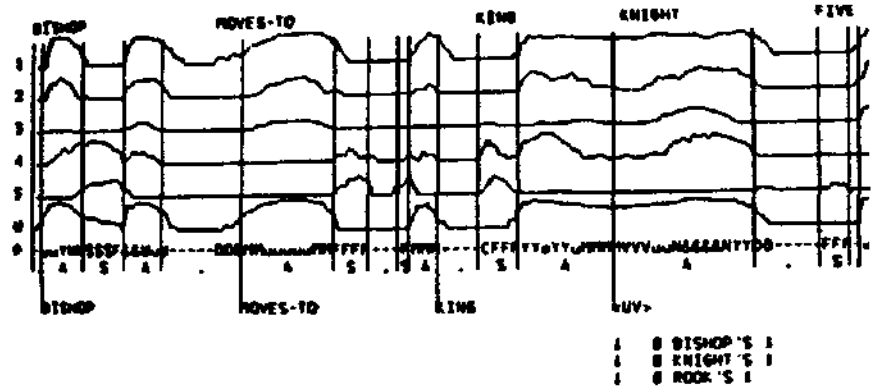


Figure 11: Seventh stage of the recognition process.

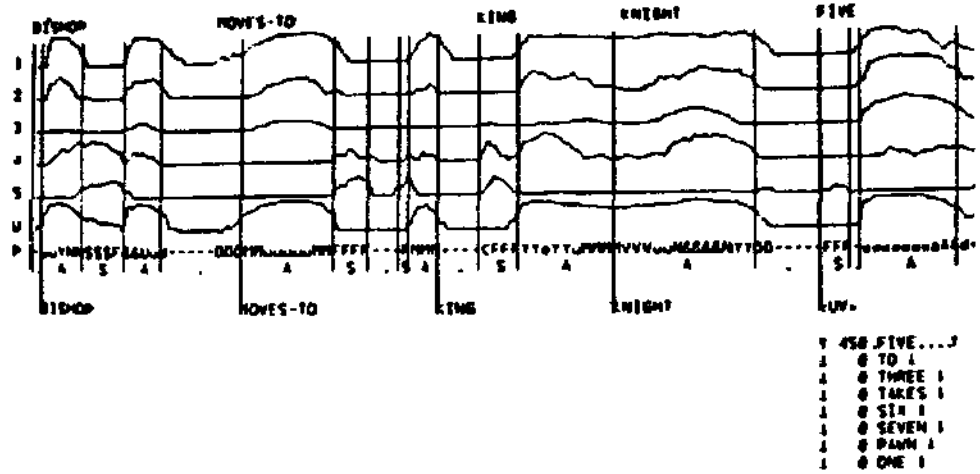


Figure 12: Eighth stage of the recognition process.

Given the top sentence hypothesis "bishop moves-to —", the Syntax module hypothesizes 7 option words. None of these ware rejected by Acoustics or Semantics. "King" and "king's" both get the highest score of 513. The first error in the recognition process occurs at this point. As new sentence hypotheses are created based on the ratings of individual words, both "bishop moves-to king's —" and "bishop moves-to king —" have the seme rating, with the former appearing at the top of the list. At this point it is instructive to see why the error was made In the first place. The phonemic description of "king's" causes a search for ■ stop followed by a vowel-like segment followed by a stop and fricative. This sequence of segments occurs in "king knight five" as can be seen from Figure 4 (improvements currently being made to the system will result in "king's" getting a much lower score). The important thing to observe is how the system recovers from errors of this type.

Figure S shows the system attempting to associate a meaningful word to the unverified part of the utterance, i.e., the /alv/ part of the word "five" in the original utterance. Syntax proposes 3 possible option words (out of a possible 31, giving a factor of 10 reduction). One is rejected and the other two get *very* low ratings. The corresponding sentence hypotheses also get low composite ratings and end up at the bottom of the stack (not visible in Figure 8).

Now we see an interesting feature of the system. In the preceding cycle (Figure 8) Syntax generated the hypotheses. It is possible that that source of knowledge is incomplete and did not generate the correct word as a possible hypothesis. Therefore, in this cycle (Figure 9), the Semantic module is given a chance to hypothesize. It hypothesizes 9 option words (a reduction of search by a factor of 3) all of which are rejected by Syntax and Acoustics. When both attempts to make a meaningful completion of the utterance fail, this particular sentence hypothesis, "bishop moves to king's--", is removed from the candidate list.

Now the top sentence hypothesis is "bishop moves-to king—" (Figure 10). Syntax hypothesizes 11 option words. Acoustics rejects six of them and Semantics rejects two. Of the remaining words, the correct word, "knight", gets the second best rating after "bishop". Again there is an errorful path, because the top sentence hypothesis now happens to be "bishop moves-to king bishop —". This sentence hypothesis is rejected immediately in the next cycle because there is no more utterance to be recognized and "bishop moves-to king bishop" is not a legal move. Note that the correct sentence hypothesis is not at the top of the stack. Its rating of 550 is not as good as "bishop moves-to king —" (see Figure 10).

The processing in the next cycle is illustrated In Figure 11. Note that in Figure 10, this same sentence hypothesis was used when the Syntax module hypothesized. Now Semantics is given an option to hypothesize and proposes 3 words. All of these are rejected by Syntax and Acoustics.

Finally, the correct partial sentence hypothesis, "bishop moves-to king knight —", gets to the top (Figure 12). Syntax hypothesizes 17 option words. Of these Semantics rejects 16 as being incorrect, leaving only "five" as a possibility. This results in the correct complete sentence hypothesis of "bishop moves-to king knight five". But the composite rating for this sentence is only 545 and there are other partial sentence hypotheses with higher ratings. At this point, the system cycles eight more times before rejecting all of them and accepting the correct sentence hypothesis.

Figure 13 shows the accuracy of the system in recognizing some typical sentences. An attempt was made to estimate the effect of syntax and semantics. Using Syntax only, the average number of words analyzed was reduced to 9.4 out of the possible 31 words in the lexicon -- a reduction In search space by a factor of 3. Using Semantics Only, the reduction of search space was about the same. Using both knowledge sources results in a reduction in the search space by a factor of 5.

```
SPOKEN
    /RECOGNIZED (If not completely correct)

pawn to queen four

pawn to queen bishop four

pawn to king four

knight to queen bishop three

bishop takes pawn

queen takes queen on queen four
    (gave up after 48 seconds of computation)

bishop to queen knight three

bishop to king three
    bishop to king five

castles queen side
    castles queen's side (understood correctly)

pawn to bishop three

pawn takes knight

knight to queen five

knight takes knight

bishop to king rook six

rook to queen three

knight to rook three

rook on rook one to queen one

rook on queen one takes rook on queen three
    rook on queen one to king rook one check

knight's pawn takes bishop

10 utterances tried:
15 recognized correctly, 16 understood correctly.  1 conceded.
Mean computation time per utterance:  10.1 sec. (PDP10 - KI10)
```

Figure 13: Examples of results for one run.

SOURCES OF KNOWLEDGE:
Their Representation and Use in the Hearsay System

Several sources of knowledge are used in the Hearsay system at present: speaker- and environment-dependent knowledge, acoustic-phonetic rules, vocabulary restrictions, and syntactic and semantic knowledge. The knowledge used at present represents only a small part of all the available knowledge. We expect to be adding to the knowledge base of the system for many years to come. The difficulties in representation and use of knowledge within the system are manifold. Even when rules exist which express pertinent knowledge, their applicability seems very limited and the effort involved to make effective use of them within the system is very large. Rules that exist are scattered in the literature. Many have not been written down and exist only in the heads of some scientists, and many are yet to be discovered. In this section, we will restrict ourselves to the discussion of the knowledge that is incorporated into the present Hearsay system.

Speaker and Environment Dependent Knowledge

The characteristics of speech vary, depending on the speaker, age, sex, and physical condition. In addition, the

characteristics of the environment (such as background noise) and the characteristics of the transducer (such as the frequency response characteristics *of* the microphone) also cause variability in speech characteristics.

In the Hearsay system an attempt is made to correct for these variables through the use of a £E table. This table contains a standard set of parameters for various phones uttered by the speaker in a neutral phonetic context. This set of parameters also accounts for the characteristics of the room noise and the characteristics of the microphone in that the neutral phones were uttered in the very same environment. A complete list of the clusters used and the details of the speaker and environment normalization are given in Erman (1973).

Acoustic-Phonetic Knowlege

This knowledge is used in several places within the system to perform different functions. Knowledge related to syllabic structure is used in the segmentation. For each segment, knowledge related to voicing, frication, and syllable junction (a local minimum of energy) is used to assign labels to each segment. An example of segmentation and labeling obtained by this type of knowledge is given in Figure 4.

The acoustic-phonetic knowledge is used in the recognition process in two ways: to generate hypotheses about possible words that may be present in the incoming utterance; and to reject, accept, or re-order the hypotheses generated by other sources of knowledge.

The hypothesization is based on the fact that certain sounds within an utterance, e.g., stressed vowels, sibilants, and unvoiced stops, can usually be uniquely recognized. These features of the incoming utterance can then be used as an acoustic-phonetic filter on the lexicon to hypothesize only those words that are appropriate in this acoustic context.

When the acoustic-phonetic knowledge is used to verify hypotheses, it *performs* a more thorough analysis. Given a hypothesized word, its phonetic description is located in the lexicon. This description is used to guide the search for the word by means of phoneme procedures. That is, the expected characteristics Of a given phoneme in various contexts are represented as a procedure; this procedure is activated to see if the expected features are present, and to provide a confidence rating based on the acoustic evidence. There are several increasingly more sophisticated verification procedures that can be used to verify proposed hypotheses. These sophisticated procedures are only invoked if word ambiguity exists at the preceding level.

Syntactic and Semantic Knowledge

Conventional parsing techniques are not very useful to direct the search within a speech understanding system. The recognizer must be capable of processing errorful strings containing spurious and repeated words. This implies that the parser must be capable of starting in the middle of the utterance where a word might be recognized uniquely and parse both forwards and backwards. The goal of parsing is not so much to generate a parse tree, but to predict what terminal symbol might appear to the left or to the right of a given context.

The predictive parsing for hypothesization is achieved in the Hearsay system by the use of anti-productions. Anti-productions act as a concordance for the grammar giving all the contexts for every symbol appearing in the grammar} they are generated from a BNF description of the language to be recognized. The anti-productions are used to predict words that are likely to occur following or preceding a word using only a limited context. Examples of anti-productions and their use are given by Neely (1973). The role *of the syntactic verifier* is to accept *or discard* hypotheses by using syntactic consistency checks based on the partial parse of the utterance. While the knowledge used for hypothesization and verification are the same, the representation and the mechanisms used in the hypothesization and verification are different. Figures 5 and 6 give examples of constraints provided by the syntactic knowledge during hypothesizatidn. Figure 9 illustrates its use in verification.

The semantic source of knowledge for Voice-Chess is based on the semantics of the task, the current board position, and the likelihood of ihe move. This knowledge is used to predict likely legal moves; these moves are then used in conjunction with the partially-recognized utterance to predict a word that might appear in the utterance. The same knowledge is also used to verify hypotheses generated by other sources of knowledge. Figure 9 illustrates the use of semantic knowledge to generate hypotheses. In the context of "bishop moves-to king", Semantics hypothesizes nine possible words. It hypothesizes all the words that might appear in the utterance in positions allowed by the semantic knowledge, given the partial recognition. Figure 12 shows the *use* of Semantics in the verification. Syntax hypothesizes 17 possible words. The semantic knowledge, given the partially recognized utterance "bishop moves to king knight", indicates that only "five" is legal in that context by rejecting all others.

SUMMARY

This paper reports *on* research in progress on the Hearsay speech understanding system. The system has been operational since June, 1972. At present we are attempting to improve the accuracy and performance of the system by adding to and improving the knowledge base. This is being done by an analysis of errors made by the system on seven sets of data from five male speakers in four different task domains. This process of modification and improvement is expected to continue for several *years,* using increasingly complex vocabularies, syntax, *and* task environments. The Hearsay system will be used primarily as a research tool to evaluate the contributions of various sources of knowledge, as well as serving as an information processing model of speech perception,

ACKNOWLEDGE

We wish particularly to acknowledge the efforts of Bruce Lowerre who has increased the acoustic-phonetic knowledge base in Hearsay, thereby greatly improving the system's performance.

BIBLIOGRAPHY

1. Erman, LD. {1973, in preparation), An Environment and System for Machine Recognition of Connected Speech, Ph.D. Thesis, Comp. Sci. Dept., Stanford Univ., to appear as a Tech. Rep., Comp. Sci. Dept., Carnegie-Mellon Univ., Pittsburgh, Pa.

2. Fant, G. (1964), Auditory Patterns of Speech", in W. Wathen-Dunn (ed), Models for the Perception of Speech and Visual Form, MIT Press.

3. Gillogly, J.J. (1972), The TECHNOLOGY Chess Program, Artificial Intelligence, 3, 145-163.

4. Halle, M., and K. Stevens (1962), "Speech Recognition: A Model and a Program for Research", IRE Trans. Inform. Theory, IT-8, 155-159.

5. Liberman, A.M., F.S. Cooper, K.S. Harris, and P.F. MacNetilage {1962), "A Motor Theory of Speech Perception", Proc. of Speech Comm. Seminar, 2, KTH, Stockholm.

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

7. Newell, A., J. Barnett, J. Forgie, C. Green, D. Klatt, J.C.R. Licklider, J. Munson, R. Reddy, and W. Woods (1971), Final Report of a Study Group on Speech Understanding Systems, North Holland (1973).

8. Reddy, D.R., L.D. Erman, and R.B. Neely (1970), The C-MU Speech Recognition Project, Proc. IEEE System Sciences and Cybernetics Conf., Pittsburgh, Pa.

9. Reddy, D.R. (1971), Speech Recognition: Prospects for the Seventies, Proc. IFIP 1971, Ljubljana, Yugoslavia, Invited paper section, pp. 1-5 to 1-13.

10. Reddy, D.R., L.D. Erman, and R.8. Neely, et al. (1972), Working Papers in Speech Recognition, Tech. Rep., Comp. Sci. Dept., Carnegie-Mellon Univ., Pittsburgh, Pa.

11. Reddy, D.R., L.D. Erman, and R.B. Neely (1972a), A Model and A System for Machine Recognition of Speech, (to be published in IEEE Trans, on Audio and Electro-acoustics, 1973).