EYES AND EARS FOR COMPUTERS

D. Raj Reddy
Computer Science Department
Carnegie-Mellon University
March, 1973

This paper presents a unified view of the research in machine perception of speech and vision in the hope that a clear appreciation of similarities and differences may lead to better information-processing models of perception. Various factors that affect the feasibility and performance of perception systems are discussed. To illustrate the current state of the art in machine perception, examples are chosen from the HEARSAY speech understanding system and the image processing portion of the SYNAPS neural modelling system. Some unsolved problems in a few key areas are presented.

Keynote speech to be presented at the Conference on Cognitive Processes and Artificial Intelligence, Hamburg, April, 1973.

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 Air Force Office of Scientific Research, and in part by the National Science Foundation under contract No. GJ32784.

EYES AND EARS FOR COMPUTERS

> It is clear that all the (visual and speech) phenomena occur in both space and time. In visual signs it is the spacial dimension which takes priority, whereas the temporal dimension takes priority in auditory signs...what is the substantial difference between spacial and auditory signs? We observe a strong tendency to reify visual signs, to connect them with objects, to ascribe mimesis to such signs, and to view them as elements of an "imitative art". . . . . On the other hand verbal and musical signs show us two essential features. First, both music and language present a consistantly heirarchized structure, and, second, both are resolvable into ultimate, discrete, rigorously patterned components which, as such, have no existence in nature but are built ad hoc.
>
> One should not draw the frequently suggested but over-simplified conclusion that speech displays a purely linear character or that visual perception is performed by purely simultaneous synthesis. Luria shows that in our perception of a painting, we first deploy step-by-step efforts to go over from certain selected details from parts to the whole, and for the contemplator of a painting the integration follows as a further phase, as a goal. In the fifth century, Bhartrhari, the great master of Indic linguistic theory, distinguished three stages in a speech event, conceptualization, production and audition, and comprehension. While production and audition are naturally sequential, both conceptualization and comprehension of the whole message is done at one and the same time. This conception is akin to the modern psychological problem of "short-term memory".
>
> Jakobson (1964)

INTRODUCTION

Visual and speech perception tasks, which can be performed with no apparent effort by people, have proved to be difficult for machines. This may be in part due to the absence of cognitive models of perception of the type proposed above by Jakobson. In this paper we attempt to give a unified view of the research in machine perception of speech and vision in the hope that a clear appreciation of similarities and differences may lead to better information processing models of perception. Being active in research in both computer vision and speech, we have found it useful to look at the problems that have arisen in one domain and anticipate corresponding problems in the other (Reddy, 1969). Thus, this paper represents a comparitive study of the issues, systems, and unsolved problems that are of interest to visual and speech recognition research at present.

To distinguish from the multitude of activities that come under the all encompassing term pattern recognition (digit recognition, isolated word recognition, character recognition, etc.),* this paper will be restricted to the areas of research denoted by "speech understanding systems" and "scene analysis". These terms represent attempts at machine perception of unrestricted speech and visual stimuli, e.g., spontaneous (possibly ungrammatical) connected speech from many speakers, and naturally occuring scenes such as people, rooms, trees etc The main problem here is not one of categorization and classification but rather that of analysis and description (Narasimhan, 1966). These areas are further characterized by the notion that, to equal human performance, many sources of knowledge (possibly disjoint) have to be brought to bear on the perception task. It is also assumed that these sources of knowledge ("capsules of intelligence") must effectively cooperate with each other to achieve better perception than would be the case when some of the sources of knowledge are absent. It is appropriate to quote Newell, et al., (1971) on this subject:

> We call the type of system to be investigated a **speech-understanding** system. The inclusion of **understanding** is to distinguish the systems somewhat from speech **recognition** systems. It does not so much indicate enhanced intellectual status, but emphasizes that the system is to perform some task making use of speech. **Thus, the errors that count are not errors in speech recognition, but errors in task accomplishment.** If the system can guess (infer, deduce,...) correctly what the user wants, then its inability to determine exactly what the user said should not be held against it — even as for you and I.

Though the eventual goal of speech understanding and scene analysis research is to accept unrestricted stimuli, we do not at present know how to design such systems. It is natural, then, to attempt to build systems which perform restricted perception tasks, e.g., recognition of isolated words from a single cooperative speaker or of a visual scene containing only rectangular parallelepipeds of different sizes and colors. However, unless these systems are designed with the eventual goal in mind, it is possible to end up with systems which are too specialized and unextendable. Thus it becomes necessary to have a global view of the problem and the many dimensions along which systems can vary. This would be helpful in designing experiments and systems which, though restricted, can provide valuable knowledge towards the ultimate system. Some of these dimensions were discussed by Reddy (1969). A more complete list for speech was given by Newell et al., (1971). In the next section we will discuss these variables in greater detail.

Although there have been many papers on the subject of speech recognition, there have only been a few working systems for the recognition of connected speech. The system of Fry and Denes (1969) was hardwired and used probabilistic

* An earlier paper by David and Self ridge (1962) also titled "Eyes and Ears for Computers" provides a summary of the level of achievement in character recognition and digit recognition systems. It is interesting to note the change in the level of expectation and aspiration within a decade as evidenced by these two papers with the same title.

information to improve recognition. The system of Sakai and Doshita (1963) was hardwired to perform segmentation, phone and word recognition. Hughes and Hemdal (1965) used a computer-based feature extraction system for the recognition of vowels and some consonants. Reddy (1967) analyzed a limited set of connected speech utterances to formulate algorithms for segmentation, phoneme grouping, and classification for many phonemes of English. The Vicens-Reddy system demonstrated the use of syntactic information (Vicens, 1969) in speech recognition. The system of Tappert and Dixon (1972) uses sequential decoding techniques in the analysis of connected speech. The Hearsay system (Reddy, Erman and Neely, 1972) is the first working connected speech recognition system using non-trivial syntax and semantics. We will describe the structure of this system in greater detail in a later section. We can expect interesting results in this area over the next several years because there are several groups active at present in speech understanding research (Barnett, 1972; Fant, 1970; Forgie, 1972; Walker, 1972; Woods, 1972). In addition, there is a great deal of relevant research in the areas of speech analysis, synthesis, and perception (Fant, 1960; Flanagan, 1965) and in the area of phonetics and linguistics (Lehiste, 1967; Chomsky and Halle, 1968).

The work in scene analysis has been centered mainly around robotics research at several artificial intelligence centers: Stanford, MIT, SRI, and Edinburgh. As such it has often been overly restrictive in scope. The papers by Feldman, et al., (1969 and 1971), Nilsson (1969), Fikes and Nilsson (1971), illustrate the state of the art in this area. Most of this work has produced a repertory of techniques for specific tasks, e.g., plane bounded convex objects, rooms without clutter, or, in general, subproblems whose main motivation is that they can be analyzed without too much difficulty or too many errors within the present state-of-the-art. However, there have been several advances: scene analysis by classification of types of intersections (Guzman, 1968), the use of the notion of planning in picture processing (Kelly, 1970), accommodation in computer vision (Tanenbaum, 1971), building structural descriptions from examples (Winston, 1971), analysis of curved objects (Krakauer, 1971), and so on (see Rosenfeld (1969, 1973) for a more complete survey).

FACTORS AFFECTING THE FEASIBILITY AND PERFORMANCE OF A PERCEPTION TASK

> Is speech input to computer possible? The question is not well posed. It depends on many things. Consider [the long list of options]. It seems annoyingly long. But each of the concerns is an essentially independent specification that, even with present knowledge, has a strong effect on the feasibility and performance of any proposed speech recognition system. Down towards the low performance end there are combinations that are not only feasible, but are beginning to be commercially advertised (e.g., "voice-button" systems). Up towards the high end the responsible posture is that only after other intermediate steps have been accomplished successfully should an estimate be made.
>
> Newell et al. (1971)

The comments of Newell et al. on speech understanding systems hold for computer vision as well. The number of factors that affect the feasibility and performance are too numerous and are likely to grow as we understand the problems better. These factors can be grouped together into several general categories: characteristics of the source, environment, receiver (transducer), sources of knowledge, performance requirements, and computing system. In this section we will examine each of these categories and the factors influencing feasibility within each.

CHARACTERISTICS OF THE SOURCE

The factors influencing the characteristics of the sources are the composition of the stimulus, variability within the stimulus, and selectability and adaptability of the stimulus. Table 1 shows the possible choices for each of these factors.

Factor	Speech	Vision
Composition of the stimulus	Isolated Words? Connected Speech?	Single objects? Many (possibly occluded) objects?
Variability of the stimulus	One speaker? Many speakers? Open population? Male? Female? Child?	Variable size? Variable color? Variable texture?
Selectability of the stimulus	Carefully selected words? Slightly selected? Free?	Carefully selected objects? Slightly selected? Free?
Adaptability of the stimulus	Cooperative speaker? Casual speaker? Playful speaker? Trained speaker? Untrained speaker?	Carefully constructed scenes Degenerate views? Impossible objects (Escher-type)?

Table I. Factors influencing the characteristics of the Source.

Composition of the stimulus

Systems for recognition of a small set of isolated words (objects) already exist. However, when the number of words (objects) gets large or the inventory contains similar words (objects) the system performance begins to degrade significantly.

Unrestricted connected speech understanding (arbitrarily complex scene analysis) is beyond the present state of the art. The main problem here is that, depending on the context, characteristics of individual words (objects) change significantly. This may be due to coarticulation (shadows), relaxed speech (occlusion), or word boundary ambiguity (object boundary ambiguity).

Variability of the stimulus

Characteristics of a given word (object) vary depending on the speaker, sex, and physical condition (size, color, and texture). If the purpose of the perception task is to identify the word (object) independent of these variables, the system must have facilities for variability normalization. Existing systems have some variability normalization but no general schemes have emerged yet.

Selectability of the stimulus

If the words (objects) to be recognized can be preselected so as to cause minimum ambiguity resulting from similarity of structure, then the system performance can be significantly improved. While this is a useful gimmick to produce economical systems, this type of preselection can lead to unextendable systems.

Adaptability of the stimulus

If the speaker can be trained and is cooperative (if a scene composition can be carefully controlled) the system sophistication can be substantially lower than if the system has to understand casual or even playful speakers (impossible objects). However natural speech (scenes) tend to be not well-formed and cannot be carefully controlled. This type of a restriction is unlikely to be useful, if the long term goal is to recognize natural speech (or scenes).

CHARACTERISTICS OF THE ENVIRONMENT

There are two factors influencing the signal quality that are independent of the source or the receiver. These are external sources of noise and the distance between the source and the receiver. Table II shows the possible causes affecting each of these.

Noise of various forms affects the reliability of analysis. Whether a given system is useable or not depends on the environment it has to operate in. A measure of robustness of a system is how it compares with the corresponding degradation in human performance under similar noise conditions.

A microphone held too close to the lips also records the lip opening before the beginning of the utterance and the expiration at the end giving the illusion of extra sounds. When held too far, there is a loss of resolution of the signal and a decreased

signal-to-noise ratio. An object too close to the camera exhibits perspective distortion and an object too far results in the loss of resolution. There is nothing much to be done except be aware and correct for the location appropriately. Note that the human being has similar limitations as well.

Factor	Speech	Vision
Noise	Airconditioning Noise? Teletype noise? Room reverberation? Hmm, haa, and cough? Cocktail party?	Flare? Out of focus? In the shadow? Cluttered view?
Distance between source and receiver	Very close? Very far?	Very close? Very far?

Table II. Characteristics of the Environment

CHARACTERISTICS OF THE RECEIVER

There are several factors associated with the transducer that affect the performance of the system. These refer to the frequency response, amplitude response, adaptation and accomodation, and other special features. Table III indicates some of the options to be considered in the design of a system.

The choice of the transducer, microphone or telephone (Vidicon or Image dissector) depends on the application, the characteristics of the digitizer (ADC), sampling rate, and so on.

To equal human performance, the microphone should have a frequency response of 50Hz-20KHz. This implies that not only should the microphone have satisfactory frequency response in that range but the analog-to-digital conversion should be at twice the rate of frequency response desired (Nyquist rate). In practice, however, it is usually adequate to digitize speech at a rate of 20,000 samples per second for a frequency response of less than 10 KHz. Further, to avoid aliasing, it is necessary to low pass filter the data so as to remove the frequency components in the signal above the frequency response. In applications where other sources of knowledge are available to compensate for the limitations of the transducer, a much more restricted frequency response may be tolerable, e.g., telephone quality response of 300 Hz-3KHz. The lower frequency response systems have difficulty disambiguating among the fricatives /f/, /θ/, and /s/.

The transducer for visual input is usually a Vidicon TV camera, an Image dissector, or a facsimile scanner. Which one is used depends on the tradeoffs within the system: real time response, accuracy of digitization, and characteristics of the stimulus, e.g., moving vs. stationary, live input vs. photograph.

There are two types of frequency responses of interest for a visual input device. One is its response to different colors, i.e., different wavelengths in the electromagnetic spectrum. The other is its response to various spacial frequencies, i.e., the smallest resolvable object within the visual field. Within the narrow fovial region the human being is able to detect objects that subtend an angle of no more than 20" of arc on the retina. Visual input systems tend to have substantially lower resolution than that unless one uses high resolution facsimile scanners.

Factor	Speech	Vision
Transducer	Microphone? Telephone?	Vidicon camera? Image dissector?
Frequency response	50-20KHz? 300-3KHz?	10†15Hz(400nm to 700nm)? Smallest resolvable object?
Sampling rate	6000per sec.? 10000? 20000?	256x256per frame? 512x512? 1024x1024? 4096x4096?
Dynamic range	20db? 40db? 60db? 80db? 100db?	
Adaptation and accomodation	Phase-locking to a conversation? Speaker normalization? Noise normalization?	Pan? Tilt? Zoom? Focus? Automatic gain conrol?
Special Features	Pitch extractor? Phase extractor? Timbre extractor?	Color detectors? Texture detectors?

Table III. Characteristics of the Transducer

The dynamic range of the system is probably the next most important factor affecting accuracy of the system. To equal human performance, the speech and vision transducer system must have at least a 100 db dynamic range (or 10†5 different resolvable levels of sound pressure or light intensity). This requires an 18 bit analog-to-digital converter (17 bits for vision since the values are all positive). For most practical purposes a 40 to 60db dynamic range is adequate, requiring 8 to 11 bits of resolution. For low dynamic range systems (20db), a 4 bit converter may be adequate.

Other factors affecting the system that are associated with the transducer are adaptation and accommodation. The human visual system is a classic example of the types of adaptation that may be useful. Not only does the pupil focus, expand, and contract depending on the brightness and depth of the field of view, but also there is an automatic guidance system for controlling the ballistic eye movements. In machine

input systems correspondingly useful features are automatic pan, tilt, zoom, and focus mechanisms. These facilities exist on some of the current systems. For speech, corresponding facilities might also be useful for speaker, noise, and transducer normalization. At present some of the normalizaton is achieved by enhancement of high frequency components of the signal. But this is probably too primitive and too little.

Other feature extractors for measuring pitch, phase, and timbre characteristics of speech, and color and textures parameters in vision have been proposed but have not been used effectively in any speech understanding or scene analysis systems to date.

CHARACTERISTICS OF SOURCES OF KNOWLEDGE

For a given task, there are usually several sources of knowledge which can significantly enhance the performance of the system. These are usually related to the structure, number, and the interrelationship among entities (words or objects) that may appear in a given scene. The structure and interrelationships can be represented in many different ways leading to different interpretations. Table IV gives some of the main sources of knowledge available for speech and visual perception tasks.

Factor	Speech	Vision
Structure of entities	No. of different phonemes? Valid sequences of phonemes? Effect of context?	No. of different shapes that make up the objects? Strong and weak structural cues?
No. of entities	Few (<100)? Many (<1000)? Unrestricted?	
Probabalistic knowledge	Frequency counts?	Digram and trigram frequencies?
Syntactic knowledge	Fixed phrases? Artificial languages? Free English?	Fixed scenes? Restricted scenes? Naturally occurring scenes?
Semantic knowledge	Task-dependent? Analysis dependent?	User and action dependent?

Table IV. Characteristics of Sources of Knowledge

In speech, words can be further decomposed into morphemes, syllables,

phonemes, and features. For any given language, one can formulate rules governing the phonological structure of the words, e.g., number of different phonemes, restrictions on sequences of phonemes, effect of context on the articulation of a phoneme, digram frequencies, etc. In addition, for a given task the vocabulary used is usually constrained. This constraint may take one of two forms -- increasing the probability of occurance of words frequently used in that task, and, secondly, declaring (arbitrarily) that only a given subset of words may be used in the sentences for this task. This second constraint, when present, further restricts the phonological rules of the language. As the number of words in the language increases, the complexity of the perception task increases. This increase in complexity is so great that there are no systems at present that can recognize vocabularies of a thousand or greater. Part of this is due to the fact that as the vocabulary increases the number of acoustically ambiguous words may also increase, e.g., "sit", "slit", "spit", "split", etc.

In vision, unfortunately, there is no well-defined structure, akin to morphemes, syllables, phonemes, etc., that characterizes the objects to be recognized. Surfaces and shape of surfaces that make up the object is probably the closest thing. However, as in the case of speech, a given task can provide restrictions about the number and structure of the objects that might appear in a scene. These restrictions might be probabalistic or ad hoc. Ambiguity in object perception might result if two different objects can produce the same profiles from different points of view, e.g., a cube viewed from the side would show a square profile; so would a pyramid when viewed from the bottom.

Both syntactic and semantic sources of knowledge primarily reflect the interrelationships affecting the composition of a sentence (a scene). The sentence (scene) composition may be arbitrarily constrained to minimize the problems of analysis. The problems that arise in speech at this level are word boundary ambiguities (see next section for some examples), changes in segmental and suprasegmental characteristics depending on sentence context, and non-grammaticality and non-well-formedness of sentences in spoken language. In vision, the problems are determining object boundaries in the presence of shadows, occlusions, and matched surface junctures. Availability of syntactic and semantic sources of knowledge of the type listed in Table IV helps to direct and focus the search during the perception task. We will see more on the use of syntax and semantics in the next section.

CHARACTERISTICS OF THE SYSTEM

There are several characteristics of the system which have by far the largest impact on the success or failure of a perception task, viz., the model (method of solution), the system organization, the desired performance, and the computing system used. Table V gives the choices available in each of these dimensions.

While a hierarchical structure may be adequate for simple recognition tasks, it is not adequate for systems which have to use many diverse sources of knowledge. Analysis-by-Synthesis (Stevens and Halle, 1961) and Heterarchical Systems (Minsky

and Pappert, 1972) are adequate but are either computationally expensive or do not lend themselves to systems organizations that satisfy the following requirements which we think important:

1. Contributions of syntax, semantics, context, and other sources of knowledge towards analysis should be clearly evaluatable. Exactly what and how much does each contribute towards improving the performance of the system?
2. The absence of one or more sources of knowledge should not have a crippling effect on the performance of the model.
3. When more than one source of knowledge is available, interactions between them should lead to a greater improvement in performance than is possible to attain by the use of any subset of sources of knowledge.
4. Since the decoding process is errorful at every stage, the model must permit graceful error recovery.
5. Increases in performance requirements (such as the real time requirement, increase in vocabulary, modifications to the syntax, or changes in semantic interpretation) should not require major reformulation of the model.

We have arrived at a model which is intended to satisfy the above requirements. It consists of a small set of cooperating independent processes capable of helping in the decoding process either individually or collectively and using the "hypothesize-and-test" paradigm. We will see the use of this model in the next section.

Factor	Options
Model	Hierarchical? Heterarchical? Hypothesize-and-test? Analysis-by-Synthesis?
System Organization	Simple program? Multiprocessing? Parallel processing? Pipeline? Feedback? Feed forward? Backtrack? Planning?
Performance	Real time? About real time? No hurry? No errors (<.1%)? Few errors (<)? Many errors (<20%)?
Processing power of computer	1 million instructions/sec? 10 mips? 100 mips? 1000 mips?
Size of memory	1 megabit? 10 mb? 100 mb? 1000 mb?
Cost (per second of speech or per scene) analyzed	.001$/s? .01$/s? .1$/s? 1.00$/s? 100.0$/s?

Table V. Characteristics of the System

System Organization

Even when the model (the method of solution) is specified, there are several

possible ways the program structure could be organized: as a single processor system, a multiprocessor system or a parallel processing system. Nonrestricted speech understanding and scene analysis tasks will probably require systems of substantially greater computational power than can be obtained by, say, a million-instruction-per-second computer. One way to achieve this power is to use several processors. Multiple processors may, in turn, require substantial reformulation of the problem solution. In addition, if each source of knowledge is to be activated as an independent process as in the model above, then the system can be programmed to run under a single processor or multiprocessor system.

A feature that characterizes all machine perception systems is that, at each stage of the processing, they make some errors while correcting others. This errorful nature of systems makes it imperative that they be fail-soft, i.e., they make no irrevocable decisions. This is achieved within a program using techniques such as backtracking, feed-forward, feed-back, etc.

Performance Requirements

The real-time requirement is probably the most difficult requirement to satisfy. To equal human performance, a system must sometimes be able to answer questions (detect motions) even before they are completed. This means that various subprocesses within the system (representing various sources of knowledge) must be able to operate on the incoming data as soon as a meaningful "chunk" of data is available without waiting for the completion of the utterance. This poses serious problems for system organization in the activation, control, and interprocess communication of the subprocesses.

The other performance requirement, accuracy, is equally demanding. By now it is axiomatic that almost any reasonable strategy for analysis will achieve 80% accuracy. Attempts at improving the performance seems to require exponentially increasing effort. The higher the accuracy requirement, the greater the tradeoffs with respect to all the other dimensions: vocabulary size (number of objects), number of speakers (colors and textures), time for analysis, and so on.

Economics of recognition

Ultimately, whether a speech understanding (scene analysis) system is used in an application or not depends on the cost of recognition. The cost in turn depends on the speed and memory requirements of the computer used in the perception task. It seems possible to build adaptive isolated word recognizers of about a 100 word vocabulary for a few thousand dollars. However, connected speech understanding systems with large vocabularies are likely to be very expensive and uneconomical for most tasks. Similarly, a simple vision system controlling a simple assembly task can be produced economically today. General purpose computer vision seems far away.

SYSTEMS FOR MACHINE PERCEPTION

> Automatic speech recognition « as the human accomplishes it will probably be possible only through the proper analysis and application of grammatical, contextual, and semantic constraints. This approach also presumes an acoustic analysis which preserves the same information that the human transducer (i.e., the ear) does. It is clear, too, that for a given accuracy of recognition, a trade can be made between the necessary linguistic constraints, and complexity of the vocabulary, and the number of speakers.
>
> J. L Flanagan (1965)

A main focus of this paper is to suggest that, to equal human performance in perception, machines must use all the available sources of knowledge. These sources of knowledge tend to be too diverse and disjoint to be used in a uniform manner. Further, some of the sources of knowledge may be absent. Thus, the absence of useful syntax and semantics in a given task should not have a crippling effect on the performance of the system. When more than one source of knowledge is available, interactions between them should lead to greater improvement in performance than is possible to attain by the use of any subset of the sources of knowledge.

Although the use of syntax, semantics, and context in a perception task have been talked about for a long time, there have been few systems which demonstrated how these sources of knowledge may be used in a the recognition task. The focus of the report by Newell, et al, (1971) was, therefore, to propose a program for research for a class of systems in which the effect of these diverse sources of knowledge could be examined.

Rather than talk about possible organizations of systems which use many sources of knowledge, we will attempt to illustrate the point by means of two specific examples: The HEARSAY System (Reddy, Erman, Fennell, and Neely, 1973) for a speech understanding task and the image processing part of the SYNAPS System (Reddy, Davis, Ohlander and Bihary, 1972a) performing a visual perception task. These systems were chosen because they are illustrative of the current state of the art and, more importantly, they are the systems most familiar to the author.

THE HEARSAY SYSTEM

HEARSAY is a speech understanding system presently under development at Carnegie-Mellon University. It is not restricted to any particular recognition task. Given the syntax and the vocabulary of a language and the semantics of a task, HEARSAY attempts recognition of the utterance in that language. 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 1 illustrates the board position for this example 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. As illustrated in Figure 2, this is one of 46 different legal moves in this position. These moves have been ordered on the basis of their goodness in the given board position. The negative rates indicate that it would be a very bad move. This judgement was based on a task dependent source of knowledge available to program (Gillogly, 1972). Note that the move chosen by the speaker was only the fourth best move in that situation.

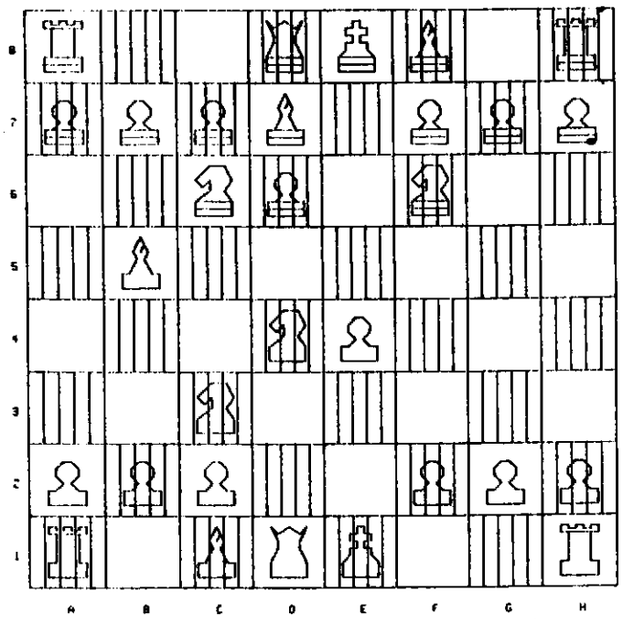


Figure 1. The chess board position at the time the move is spoken.

Legal Moves

(Ordered by ratings)

Move	Rating	Move	Rating
0-0	0	KNP/KN2-KN3	0
QB/QB1-K3	0	QNP/QN2-QN3	0
QB/QB1-KB4	0	QRP/QR2-QR3	0
QB/QB1-KN5	0	KN/Q4-K2	0
KB/QN5-QB4	0	KN/Q4-QN3	0
KR/KR1-KB1	0	QN/QB3-Q5	-100
Q/Q1-Q3	0	KP/K4-K5	-100
KB/QN5XQN/QB6	0	K/K1-K2	-100
KBP/KB2-KB4	0	QN/QB3-K2	-100
KR/KR1-KN1	0	KNP/KN2-KN4	-100
KBP/KB2-KB3	0	QNP/QN2-QN4	-100
Q/Q1-Q2	0	QN/QB3-QR4	-100
QR/QR1-QN1	0	QN/QB3-QN1	-100
K/K1-KB1	0	QB/QB1-Q2	-330
KB/QN5-K2	0	KB/QN5-Q3	-330
KN/Q4-KB5	0	Q/Q1-KB3	-330
KB/QN5-QR4	0	Q/Q1-K2	-330
KN/Q4XQN/QB6	0	QB/QB1-KR6	-330
KN/Q4-KB3	0	K/K1-Q2	-330
KRP/KR2-KR4	0	KN/Q4-K6	-330
QRP/QR2-QR4	0	KB/QN5-QR6	-330
KB/QN5-KB1	0	Q/Q1-KN4	-900
KRP/KR2-KR3	0	Q/Q1-KR5	-900

Figure 2. A list of the legal moves for the board position in Figure 1.

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> <position>. The piece can have qualifiers to indicate the location. The action may be of the form: "to", "moves-to", "goes-to", "takes", "captures", and so on. The position is of the form: "king three", "king bishop four", or "king 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 just as "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 manual segmentation and labelling 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 is repeated for the whole duration of the word. Thus, the

word "bishop" was articulated for 400 milliseconds and occupies the first two rows of the wave form. The second line of text under each row contains the phoneme being articulated. The phoneme (represented in IPA 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 phoneme /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 Segmentation

The speech input from the microphone is passed through five octive band-pass filters (spaning 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 parameters every 10 milliseconds. These parameters are smoothed and log-transformed. Figure 4 shows a plot of these parameters as a function of time. The top line of the figure indicates the location where each word of the utterance begins as marked during the manual segmentation process (this will permit us to verify the accuracy of the machine recognition process in the later stages).

This vector of parameters (labeled 1, 2, 3, 4, 5, and U in Figure 4) are compared with a standard set of parameter vectors to obtain a minimum distance classification using a nearest neighbor classification technique. The line of text labeled P in Figure 4 gives the classification for every 10 millisecond unit. The standard set of parameters is obtained by selecting parameter values from a training set of utterances containing various phonemes in neutral context. When a phoneme is represented by several articulatory jestures more than one cluster center may be added to the standard set. This technique provides a way for correcting for the characteristics of the source (speaker variations), characteristics of the environment (noise), and characteristics of the receiver (microphone) that we discussed in the previous section.

The classification of labels so obtained (row P in Figure 4) is then used to specify a feature set, such as voicing and frication, and these features are used in the segmentation of the utterance, shown in Figure 4. The boundaries of segments are

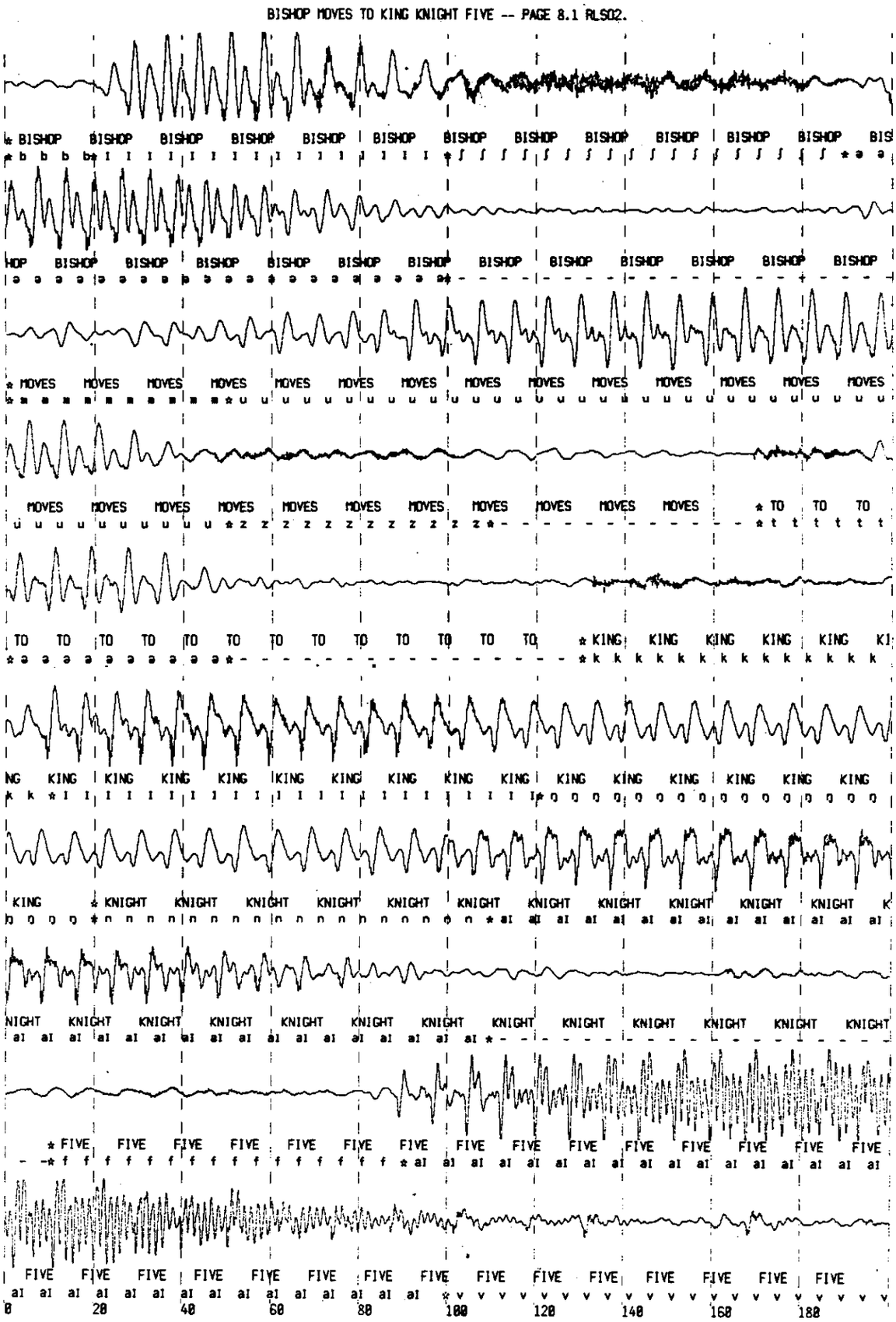


Figure 3. Waveform of the utterance showing the actual word and phoneme boundaries.

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).

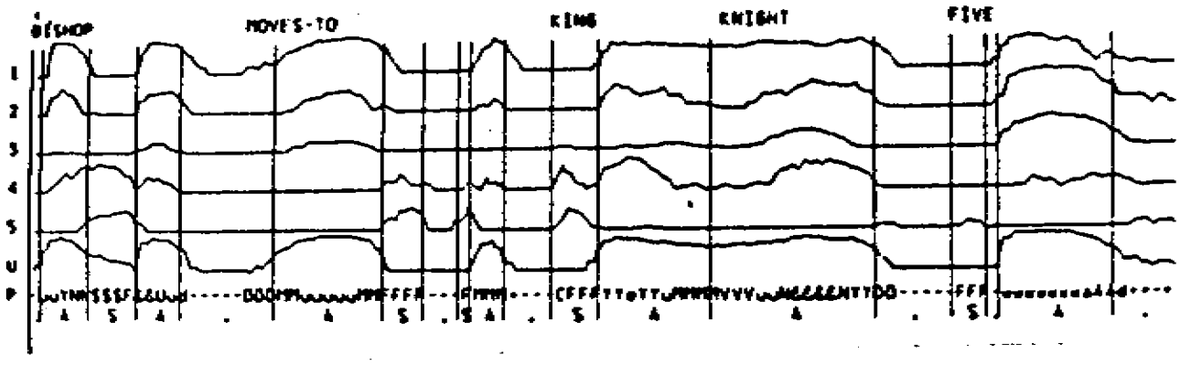


Figure 4. Parametric representation of the utterance showing the results of feature extraction and segmentation.

The Recognition Process

The HEARSAY system 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 languages 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.

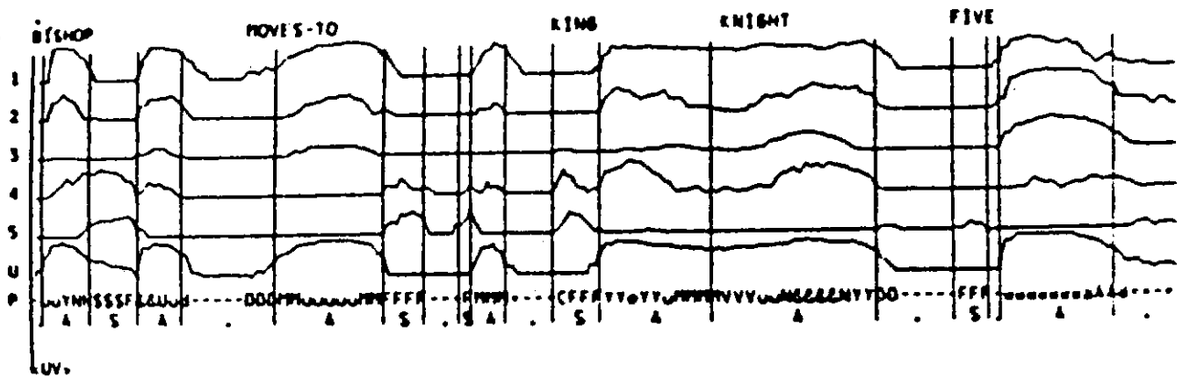
These processes cooperate by means of a hypothesize-and-test paradigm. This paradigm consists of one or more sources of knowledge looking at the unrecognized portion of the utterance and generating an ordered list of hypotheses. These hypotheses may then be verified by one or more of the sources of knowledge; 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..., ACO REJECTED...), the acceptable option words with their ratings and word boundaries (e.g., ↑...↑ 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 "↑...↑" 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 of 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 rejects the word "bishops" as being 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 top 8 words is 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.

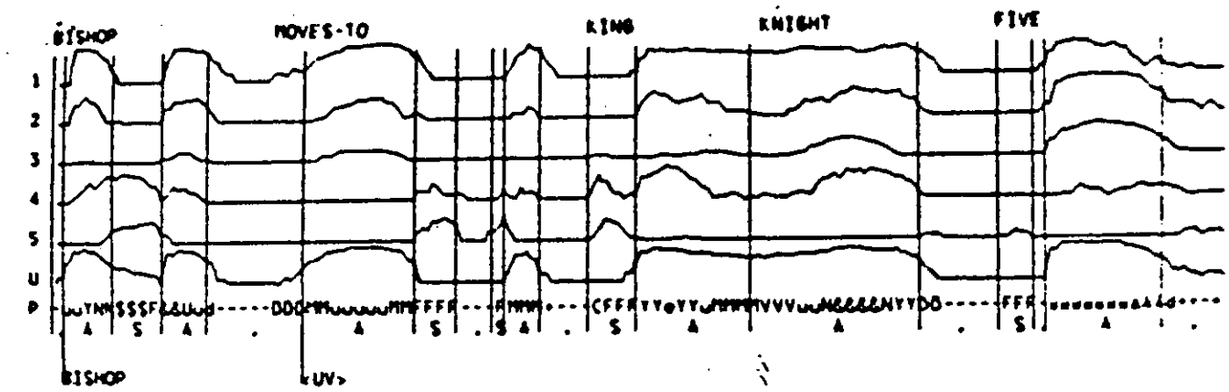
Figure 6 shows the second cycle of the recognition process. The top sentence hypothesis is "bishop ---". An attempt is being made to recognize the word following "bishop". Again Syntax generates the hypotheses. Given that "bishop" is the preceding word, the syntactic source of knowledge proposes only 7 possible 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 547.

Figure 7 shows the third cycle of the recognition process. Given the top sentence hypothesis "bishop moves-to ---", the Syntax module hypothesizes 7 option words. None of these were 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 same 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 a stop followed by a vowel-like segment



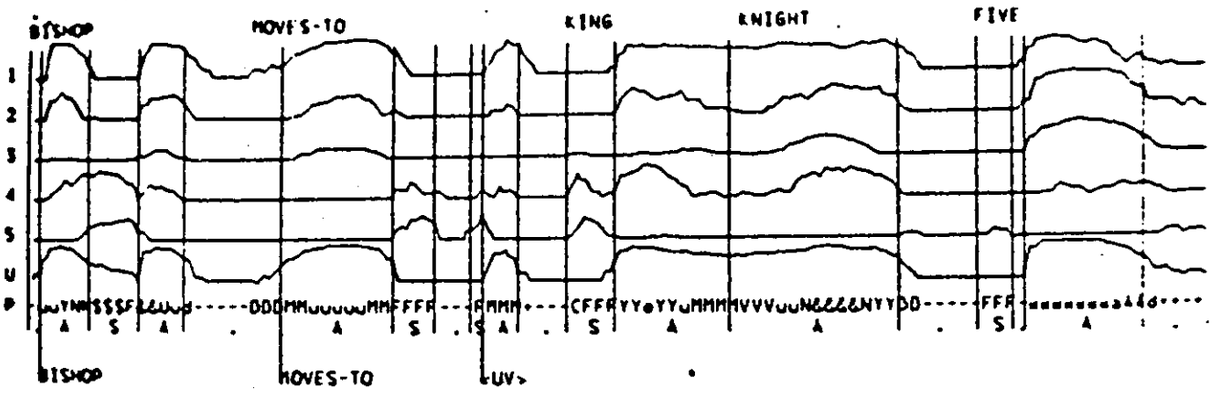

```
 ?........? 500 ROOK'S
 ?........? 500 ROOK
?.........? 500 QUEEN'S
?....? 500 QUEEN
?....? 425 PAWN
?....? 463 KING
 ? 513 BISHOP.........?
?.........? 463 KING'S

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:
( 533) BISHOP <FILLER>
( 528) ROOK'S <FILLER>
( 528) ROOK <FILLER>
( 528) QUEEN'S <FILLER>
```

Figure 5. First stage of the recognition process.



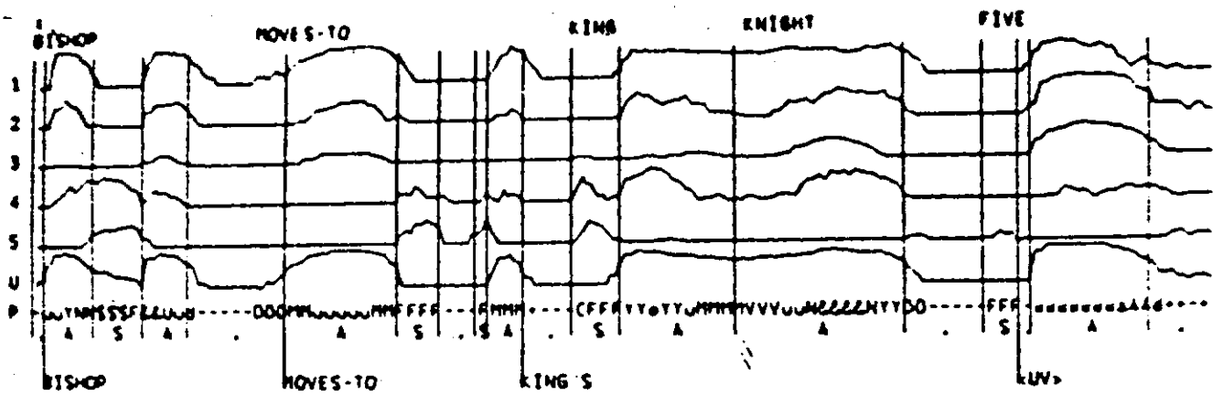

```
                ? 443 TO.........?
                ? 345 TAKES.....?
                ? 455 PAWN.....?
                        ? 435 ON.?
                        ? 525 MOVES-TO...?
                        ? 463 GOES-TO...?
                        ?   0 CAPTURES ?

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>
( 533) 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




```
                    ↑ 368 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

followed by a stop and fricative. This sequence of segments occur in "king knight five" as can be seen from Figure 4 (improvements 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 8 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 -- factor of 10 reduction). One is rejected and the other two get very low ratings. The corresponding sentence hypotheses also get a low composite rating and end up at the bottom of the stack (not visible in Figure 8).

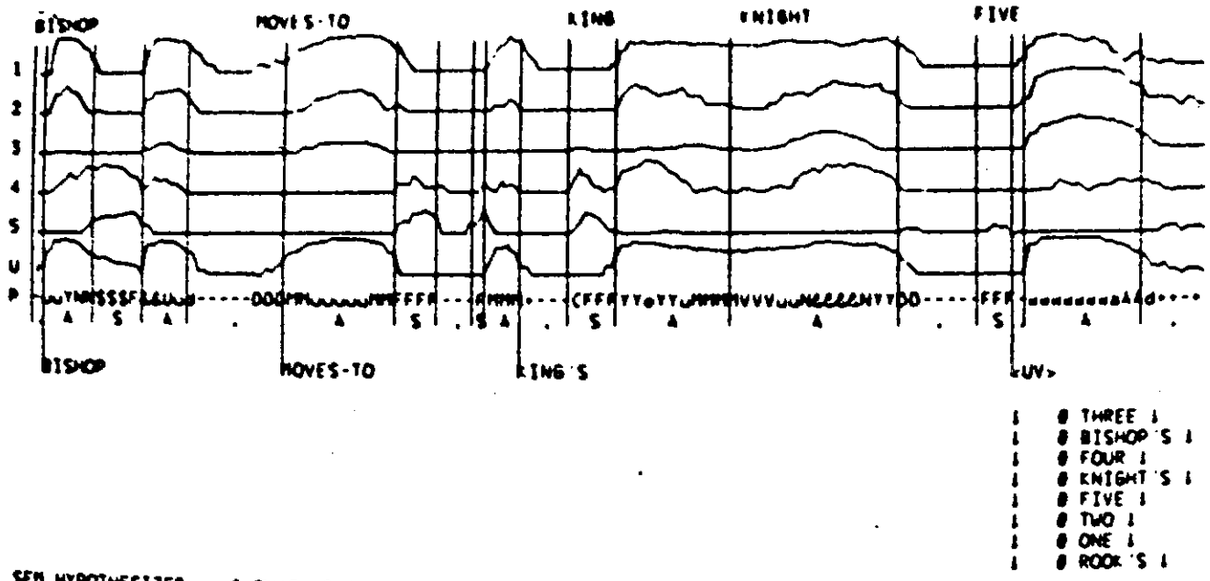
Now we see an interesting feature of the system. In the preceding cycle (Figure 8) Syntax generated the hypotheses. It is possible that the syntactic source of knowledge is incomplete and did not generate the correct word as a possible hypothesis. Therefore, in this cycle (Figure 9), the Semantic module 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 with Syntax module hypothesizing. Now Semantics is given an option to hypothesize and proposes 3 words. All of these are rejected by Syntax and Acoustics.

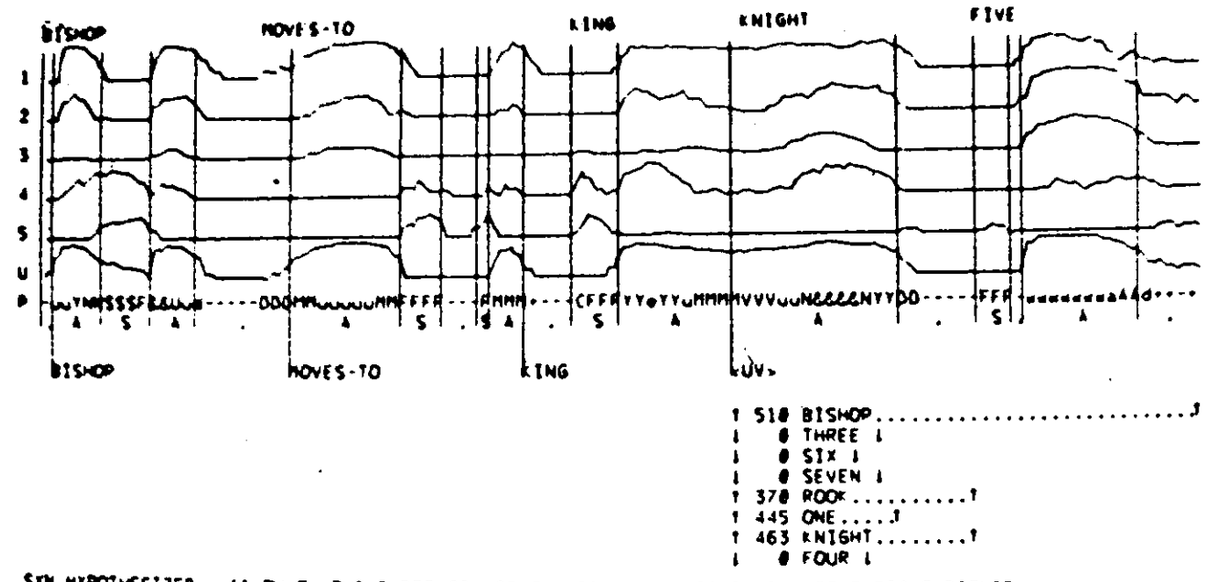
Finally, the correct sentence hypothesis, "bishop moves-to king knight ---", gets to the top (Figure 12). Syntax hypothesizes 17 option words. Of these Semantics rejects 16 as being incorrect leaving only "five" with a positive score. This results in the correct complete sentence hypothesis of "bishop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

The HEARSAY system was demonstrated with live connected speech input in June, 1972. It is the first demonstratable system to use non-trivial syntax and semantics in the recogntion process. It is obvious from the example above that various sources of knowledge aid significantly in the reduction of search space. The system is being actively modified to increase its performance, as well as to use it as



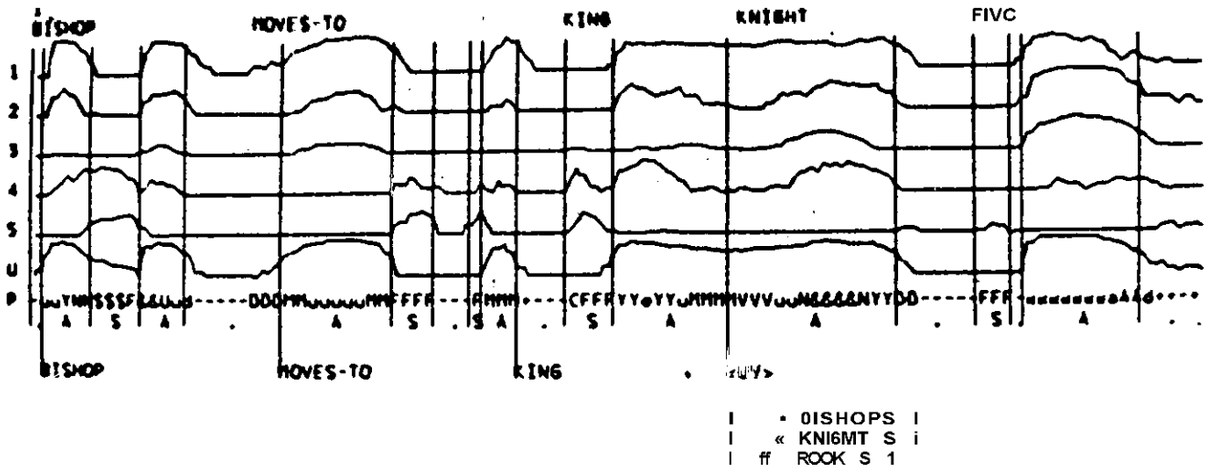

```
SEM HYPOTHESIZED     9 OW'S: THREE,BISHOP'S,FOUR,KNIGHT'S,FIVE,TWO,ONE,ROOK'S,SIX
ACO REJECTED         7 OW'S: THREE,BISHOP'S,FOUR,KNIGHT'S,FIVE,ROOK'S,SIX
SYN REJECTED         9 OW'S: THREE,BISHOP'S,FOUR,KNIGHT'S,FIVE,TWO,ONE,ROOK'S,SIX

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

Figure 9. Fifth stage of the recognition process



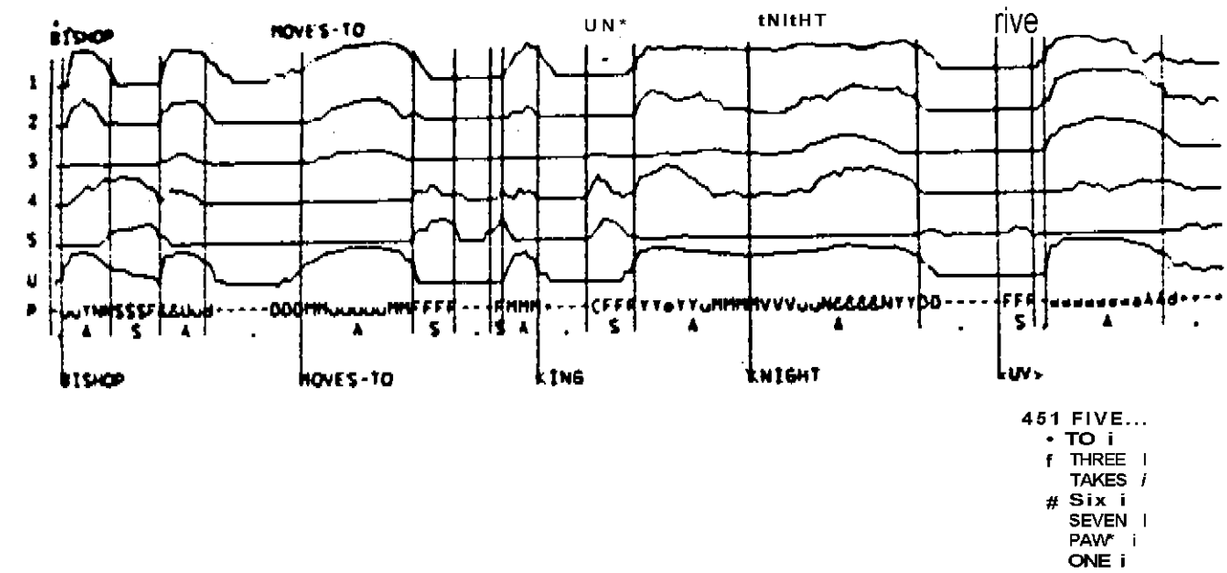

```
SYN HYPOTHESIZED    11 OW'S: TWO,THREE,SIX,SEVEN,ROOK,ONE,KNIGHT,FOUR,FIVE,EIGHT,BISHOP
ACO REJECTED         6 OW'S: TWO,THREE,SIX,SEVEN,FOUR,FIVE
SEM REJECTED         2 OW'S: SEVEN,EIGHT

THE TOP 4 SENTENCE HYPOTHESES ARE:
( 555) BISHOP MOVES-TO KING BISHOP
( 552) BISHOP MOVES-TO KING <FILLER>
( 550) BISHOP MOVES-TO KING KNIGHT <FILLER>
( 547) BISHOP MOVES-TO <FILLER>
```

Figure 10. Sixth stage of the recognition process



BISHOP S.KNI6HTS.ROOK'S
BISHOP'S
SOWS. BISHOP S.KNI6MT S.ROOK S

Figure 11. Seventh stage of the recognition process



17 OW Si TWO.TG.TMREE,TA*ES.SIX.SEVEN,PAWN.ONE.ON.hOVES-TO.GOES-TO.FOUR.HvE••
S.SI*.ONE;ON.HOVES-TO.GOES-TO.EIGHT,CAPTURES
16 Si TWO, TO. THREE .TAKES. SIX .SEVEN. PAWN .ONE .ON .HOVES-TO.60ES-TO.FOUR.

HYPOTMFSES ARE,

<F1LIER>
IISHOP IIOVES-TO QUEEN'S <F1LLER>
QUEEN <FIUER>

Figure 12. Eighth stage of the recognition process

an experimental tool for studying speech understanding, recognition, and perception. More detailed descriptions of the system are given in Reddy et al. (1972), Erman (1973), Neely (1973), Reddy et al. (1973).

THE SYNAPS SYSTEM

The SYNAPS system (Symbolic Neuronal Analysis Programming System) is being developed at Carnegie-Mellon University for the three dimensional reconstruction of dye-injected serial sections of ganglia. The eventual goal of this project is to reconstruct the complete map of neuronal connections (wiring-diagram) of a mini-brain of an invertebrate nervous system. A major component of this research is to digitize and analyze images of dye-injected histologically-prepared sections to determine locations of all dendritic structures crossing the section.

Figure 13. Photomicrograph of a section of a ganglion.

Although this is a specialized problem in image processing, the absence of well-defined boundaries and the presence of excessive noise makes it necessary to bring to bear several task-specific sources of knowledge to successfully complete the image analysis task. The purpose of the image analysis task is to extract relevant information such as the boundary of the ganglion, dendritic profiles, and other neuronal "landmarks".

The image to be analyzed is shown in Figure 13. This image is digitized using an image dissector, resulting in a matrix of light values (densities) representing the original section. Figure 14 shows a gray-scale printout of the digitized image using a Xerox Graphic Printer (Reddy, et al., 1972b). Limitations of the paper size on the XGP make it necessary to show only a coarse resolution picture of the original image.

Simple edge detection operations of the type used in earlier scene analysis programs results in the image shown in Figure 15. Note that many undesired regions appear in the output. This is to be expected given the noisy nature of the original image in Figure 13. This noise results from many sources: intensity differences caused by variable light transmission from **One** region to the next in a section, artifacts such as tissue or dust particles, unanticipated folds in the tissue, photographic distortions, uneven lighting of the microscopic field, undesired leakage from the injected neuron, etc.

There are, at the same time, several available sources of knowledge:

a. We are dealing with a known species with known landmarks which can be located uniquely from experiment to experiment.
b. The locations of desired profiles will only differ slightly from the previously analyzed adjacent section (so called continuity hypothesis).
c. Having located one profile, it is possible to extrapolate to find other profiles in the image, based on the knowledge of the corresponding profiles analyzed in the preceding section.

The effect of these sources of knowledge is to reject uncorrected spurious edges. Figure 16 illustrates the possible reduction in noise from the use of such techniques.

The SYNAPS system is still under development. The full impact of various sources of knowledge has not been evaluated yet in this system* Descriptions of non-image processing parts of the system such as 3-D reconstruction, display, and pattern analysis are given in Reddy et al (1972a).

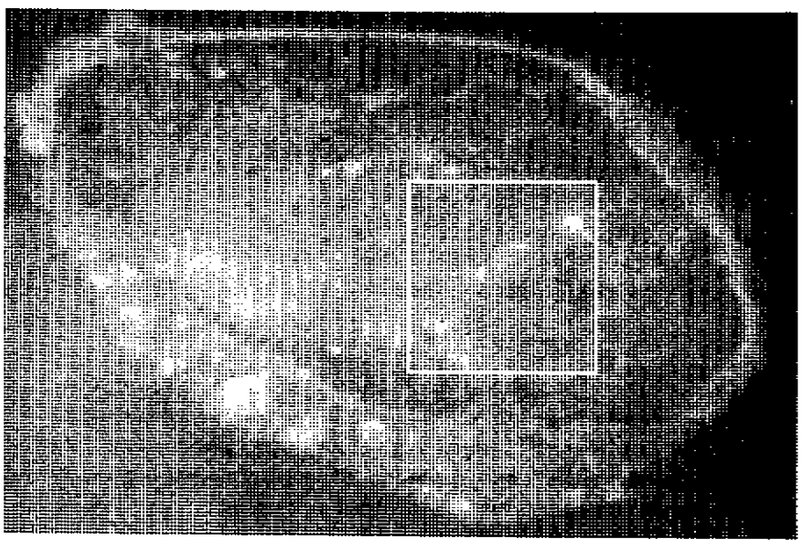


Figure 14. Gray-scale printout of the digitized image.

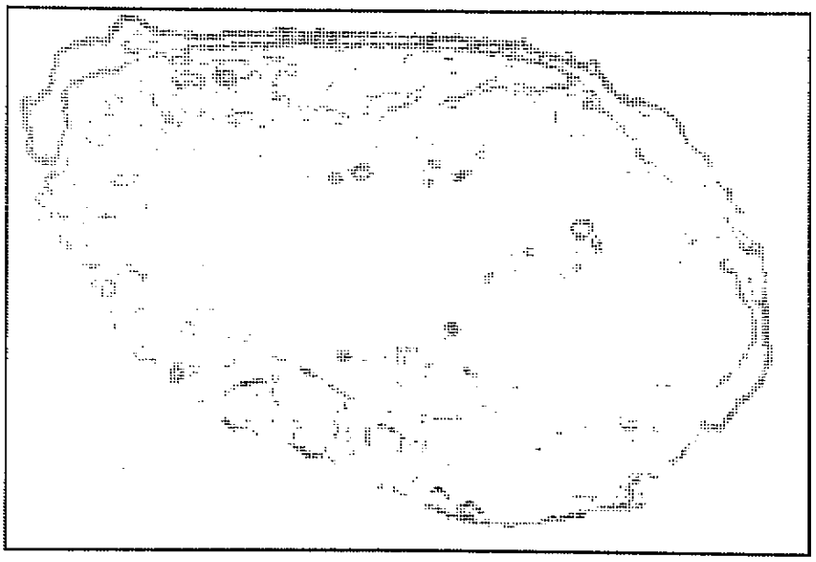


Figure 15. Result of an edge-detection operation on the digitized image.

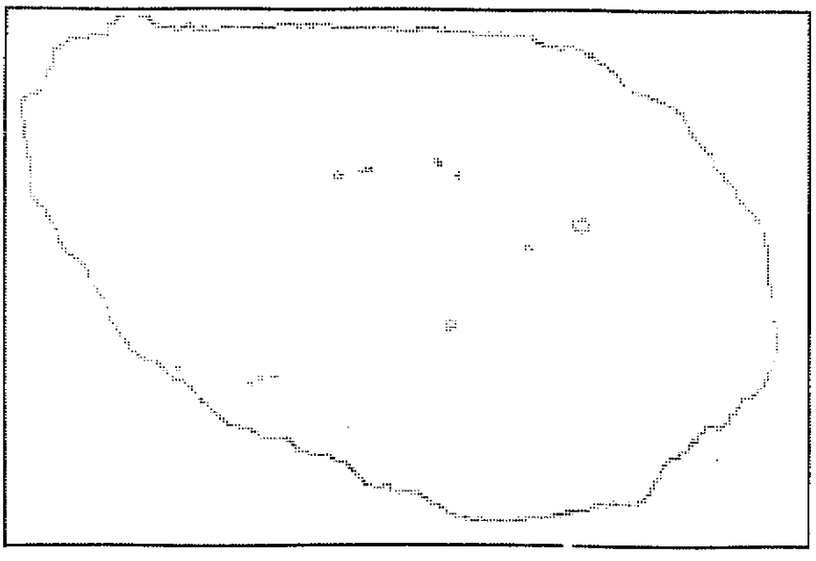


Figure 16. Noise reduction through the use of contextual information.

SOME UNSOLVED PROBLEMS

> "...lead us to believe that performance will continue to be very limited unless the recognizing device understands what is being said with something of a facility of a native speaker (that is, better than a foreigner who is fluent in the language). If this is so, should people continue work toward speech recognition by machine?"
>
> Pierce (1969)

In spite of two decades of research, progress in the fields of computer vision and speech has been very limited. When one looks for reasons for this slow and unsteady progress one finds that over-optimism, inadequate technology, and incorrect models have been the prime causes. As in the case of much of artificial intelligence research, it has proved to be difficult to build on each others' research in these areas. Significant advances in a few key problem areas could lead to rapid progress in computer vision and speech research. These problem areas can be summarized by three keywords: tools, knowledge, and theory.

Performance Evaluation

One of the features of existing perception systems, and undoubtedly of future ones as well, is the existence of error at every level of analysis and consequent proliferation of heuristic devices throughout the system to control such error and permit recycling with improved definitions of the situation. Almost entirely missing from the literature, not only of speech and vision but elsewhere in artificial intelligence as well, are techniques for evaluating performance characteristics of proposed algorithms and heuristics. By techniques, we mean both suitable instrumentation and experimental design to measure accuracy, response time, cost, etc., in relation to vocabulary, language, and context. Until such techniques are developed and applied to existing components of a perception system, these components should be considered of questionable value in an applied system.

Knowledge Aquisition

When one attempts to build a speech understanding or a scene analysis system, one finds that there are large numbers of unanswered questions. Although there have been large amounts of speech and vision research, much of it is defocused and not relevant to machine perception research. For example, there has not yet been a systematic acoustic-phonetic study of all the allophonic variations of phonemes of English. Thus it becomes necessary for the systems to "learn" acoustic-phonetic, syntactic, and semantic rules by abstraction from exemplars. We do not know how to build systems that can abstract such complex information. We will illustrate the issue by considering computer vision, but the comments are equally applicable to speech as well.

When the computer's vision system finds an object in the scene which has not been previously observed, then it seems reasonable to provide the system with the ability to question its master about the object, its structure, the utility, and the

likelihood of occurrence. If the master is unable to provide the system with an accurate description of the object (which may be often the case) then the system would have to abstract its own set of features and characteristics about this object and its relationships to the rest of the scene. This may well require several views of the object and further abstractions about the color and texture of the object.

Systems capable of building models from several views of the object have been proposed but have proved to be of limited use so far. A recent thesis by Winston (1971) attempts to abstract structural descriptions from line drawings. Abstaction from naturally occurring scenes will perhaps remain a major unsolved problem for some time to come because different objects seem to require different strategies for abstraction. Abstractions of structural descriptions of people and cars and grass and water may well require assistance from a human being before they can be effectively formulated.

This raises the issue of our ability to use partial models, both in the analysis of scenes and in acquisition of knowledge from actual views of the scene. For example, the fact that only the human hand or face is visible in a scene should be sufficient to formulate a hypothesis that the rest of the person is also attached even though he is not actually visible in the scene. Scene analysis systems must be capable of recognition of partially occluded objects where only a substructure of the object (as indicated by a partial parse perhaps) is visible. Similarly, in the acquisition of knowledge, given a partial stick figure or a caricature of an object, the computer system should be capable of abstracting the rest of the relevant characteristics from the actual scene itself.

Information Processing Models

In addition to building experimental systems for perception, we need to work on the theory of perception as well. There are many theories of perception. What we mean here are the so called **information-processing** models of perception. The notion of an information-processing model reflects a current trend in cognitive psychology to view man as an information processor, i.e., his behavior can be seen as the result of a system consisting of memories containing discrete symbols, symbolic expressions, and processes which manipulate these symbols (Newell, 1970). The main advantage of this approach to perception studies is that it permits a researcher to look at the total problem of perception at a higher functional and conceptual level than is possible with stimulus-response studies and neuro-physiological models.

There is a great deal of work in cognitive psychology on memory representations (e.g., Sperling, short-term, and long-term), on attention phenomena and serial vs. parallel processing, on EPAM-like pattern matching, and on perceptual illusions (Simon, 1967; Simon, 1972; Simon and Barenfeld, 1969; Newell and Simon, 1972; Newell et al., 1973; Chase and Simon, 1973). But much of the work in computer vision does not seem to benefit from this work. Conversely, many of specific models of machine perception, such as cooperating independent processes, utilization of sources of knowledge, hypothesize-and-test paradigm, etc., have not found their way into information processing models in cognitive psychology. This symbiosis of the two areas seems essential for significant advances in either area.

There are many other unsolved problems in machine perception(Newell et al., 1971; Montanari and Reddy, 1971). Each of the Factors discussed in the earlier section poses an unsolved problem when all the restrictive options are removed. We chose to single out the problems of tools, knowledge, and theory here because they seem to be crucial for significant advances in machine perception.

CONCLUSION

This paper has discussed issues affecting the feasibility and performance of machine perception systems, outlined the structure of the HEARSAY speech understanding system and the image analysis part of the SYNAPS neural modelling system, and posed some unsolved problems. The main focus of the paper has been to present a unified view of the research in machine perception of speech and vision.

A main question of interest is "what is the role of computer vision and speech research in artificial intelligence?". Unlike other problems in artificial intelligence, perception problems are typified by high data rates, large amounts of data, and the availability of many diverse sources of knowledge. Contrast this to many problem solving systems in which weaker and weaker methods are used to solve a problem using less and less information about the actual task. A major problem in AI, then, is develop paradigms which can effectively use all the available sources of knowledge in problem solution. Thus, the role of perception research in AI is to address itself to the questions of task representations, data representaions, and program organizations which will permit effective use of many sources of knowledge in solving problems involving high data rates and large masses of data in close to real time.

A question to be answered eventually is how the human perceptual activity differs from other aspects of intelligent behavior. This raises several questions.

1. Why is it that man is able to see and hear without any conscious effort while requiring a great deal of intellectual effort to play chess or prove a theorem?
2. Does man use significantly different mechanisms for perceptual and intellectual tasks?
3. Why is it that machines seem to have as much (or more) difficulty with perceptual tasks as they do with intellectual tasks?

The answer to these and other similar questions is "We are not sure". Before we are sure, there will have to be several breakthroughs in artificial intelligence.

ACKNOWLEDGEMENT

The author would like to thank Lee Erman, Rick Fennell, Allen Newell, and Herb Simon for their valuable comments about this paper.

REFERENCES

Barnett, J. (1972), A Vocal Data Management System, IEEE Conference on Speech Communication and Processing, Boston, 340-343.

Chase, W.G. and H.A. Simon (1973), Perception in Chess, Cognitive Psychology, 4, 55-81.

Chomsky, N. and M. Halle (1968), **The Sound Pattern of English**, Harper and Row, New York.

David, E.E. and O.G. Selfridge (1962), Eyes and Ears for Computers, Proc. of the IRE, 50, 1093-1101.

Erman, L.D. (1973), An Environment and System for Machine Recognition of Continuous Speech, Ph.D. Thesis, Stanford Univ., to appear as a Technical Report, Computer Science Dept., Carnegie-Mellon Univ., Pittsburgh, Pa.

Fant, G. (1960), **Acoustic Theory of Speech Production**, Mouton and Company: The Hague.

Fant, G. (1970), Automatic Recognition and Speech Research, Quarterly Progress Report, 16-31, Dept. of Speech Communication, KTH, Stockholm.

Feldman, J.A., et al. (1969), The Stanford Hand Eye Project, Proc. IJCAI, May 7-9, Washington, D.C.

Feldman, J.A. et al. (1971), The Use of Vision and Manipulation to Solve the "Instant Insanity" Puzzle, Proc. Second IJCAI, London, 359-365.

Fikes, R.E. and N.J. Nilsson (1971), STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving, Proc. Second IJCAI, London, 608-621.

Flanagan, J.L. (1965), **Speech Analysis, Synthesis, and Perception**, Academic Press: New York. Second edition, 1971.

Forgie, J. (1972), Personal Communication, MIT Lincoln Laboratories, Lexington, Mass.

Fry, D.B. and P.B. Denes (1959), The Design and Operation of a Mechanical Speech Recognizer, J. British IRE, 19, 211-229.

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

Hughes, G.W. and J.F. Hemdal (1965), Speech Analysis, Tech. Rept, AFCRL-65-681, Purdue Univ., Lafayette, Ind.

Jakobson, R. (1964), About the Relation between Visual and Auditory Signs, in Wathen-Dunn(ed), **Models for the Perception of Speech and Visual Form** MIT Press, Cambridge, Mass., 1-7.

Kelly, M.D. (1970), Visual Identification of People by Computers, AIM-130, Ph.D. thesis, Computer Science Dept., Stanford Univ., Stanford, Ca.

Krakauer, L.J. (1971), Computer Analysis of Visual Properties of Curved Objects, Ph.D. Thesis, Electrical Engineering Dept., MIT, Cambridge, Mass.

Lehiste, I. (1967), **Readings in Acoustic-Phonetics**, MIT Press, Cambridge, Mass.

Minsky, M. and S. Papert (1972), Artificial Intelligence, Technical Report, AI Group, MIT, Cambridge, Mass.

Montanari, U. and D.R. Reddy (1971), Computer Processing of Natural Scenes: Some Unsolved Problems, Proc. AGARD Symposium on Artificial Intelligence, Rome.

Narasimhan, R. (1966), Syntax-Directed Interpretation of Classes of Pictures, CACM, 9, 3, 166-173.

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 Technical Report, Computer Science Dept., Carnegie-Mellon Univ., Pittsburgh, Pa.

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 (to be published, 1973).

Newell, A. (1970), Remarks on the Relationship between Artificial Intelligence and Cognitive Psychology, in Banerji and Mesarovic (eds.), **Non-Numerical Problem Solving**, 363-400, Springer-Verlag.

Newell, A. and H.A. Simon (1972), **Human Problem Solving**, Prentice-Hall.

Newell, A. et al. (1973), Visualization, unpublished research, Carnegie-Mellon Univ., Pittsburgh, Pa.

Nilsson, N.J. (1969), A Mobile Automaton: An Application of Artificial Intelligence Techniques, Proc. IJCAI, May 7-9, Washington, D.C.

Pierce, J.R. (1969), Whither Speech Recognition, J. Acoust. Soc. Am. 46 1049-1051.

Reddy, D.R. (1967), Computer Recognition of Connected Speech, J. Acoust. Soc. Am., 42. 2, 329-347.

Reddy, D.R., (1969), On the Use of Environmental, Syntactic, and Probabilistic Constraints in Vision and Speech, AIM 78, Computer Science Dept., Stanford Univ., Stanford, Ca.

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

Reddy, D.R., W.J. Davis, R.B. Ohlander, and D.J. Bihary (1972a), Computer Analysis of Neuronal Structure, Technical Report, Computer Science Dept., Carnegie-Mellon Univ., Pittsburgh, Pa.

Reddy, D.R., B. Broadley, L. Erman, R. Johnsson, J. Newcomer, G. Robertson, and J. Wright (1972b), XCRIBL, A Hardcopy Scan Line Graphics System for Document Generation, Technical Report, Computer Science Dept., Carnegie-Mellon Univ., Pittsburgh, Pa.

Reddy, D.R., L.D. Erman, R. Fennell, and R.B. Neely (1973), The HEARSAY Speech Understanding System, to be published.

Rosenfeld, A. (1969), **Picture Processing by Computer**, Academic Press, N.Y.

Rosenfeld, A. (1973), Progress in Picture Processing: 1969-71. Computing Surveys 5, in press.

Sakai, T. and S. Doshita (1963), The Automatic Speech Recognition System for Conversational Sound, IEEE Trans., ED-12, 835-846.

Simon, H.A. (1967), An Information Processing Explanation of Some Perceptual Phenomena, Br. J. Psychol., 58, 1-12.

Simon, H.A. (1972), What is Visual Imagery? An Information Processing Interpretation, in L. Gregg (ed.), **Cognition in Learning and Memory**, Wiley, N.Y.

Simon, H.A., and M. Barenfeld (1969), Information Processing Analysis of Perceptual Processes in Problem Solving, Psychological Review, 76, 473-483.

Tenenbaum, J.M. (1970), Accomodation in Computer Vision, Ph.D. Thesis, Computer Science Dept., Stanford Univ., Stanford, Ca.

Vicens, P.J. (1969), Aspects of Speech Recognition by a Computer, Ph.D. Thesis, AIM 85, Computer Science Dept.ment, Stanford Univ., Stanford, Ca.

Walker, D. (1972), Personal Communication, Stanford Research Institute, Menlo Park, Ca.

Winston (1971), Learning Structural Descriptions from Visual Scenes, Ph.D. Thesis, Electrical Engineering Dept., MIT, Cambridge, Mass.

Woods, W. (1972), Personal Communication, Bolt, Beranek and Newman, Cambridge, Mass.