MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY
PRACTICABLE. THE COPY FURNISHED
TO DTIC CONTAINED A SIGNIFICANT
NUMBER OF PAGES WHICH DO NOT
REPRODUCE LEGIBLY.

MINUTES OF THE

SPEECH UNDERSTANDING WORKSHOP

convened on

13 November 1975

in

Washington, D.C.

DTIC
ELECTE
JAN 3 1983

S   D

B

TABLE OF CONTENTS

RE: Statement on Page 6, Attachment 12:
Not for Publication Unless Officially
Released.
Document is Unlimited and the statement
is no longer Valid per Ms. Motyka,
DARPA/TIO

DTIC
COPY
INSPECTED
2
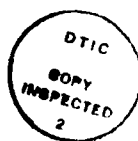
Accession For	
NTIS GRA&I	✓
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	23 cp

ii

ATTACHMENTS

MINUTES OF THE
SPEECH UNDERSTANDING WORKSHOP
CONVENED ON
13 NOVEMBER 1975
IN
WASHINGTON, D.C.

1.1   MORNING SESSION

The meeting was called to order by Major Carlstrom,
USAF, Defense Advanced Research Projects Agency, Information
Processing Techniques Office (ARPA/IPTO) at 8:45 hours.

1.2   A roster of attendees and a copy of agenda items are
appended as Attachments 1 and 2 respectively.

1.3   MAJOR CARLSTROM'S INTRODUCTORY REMARKS

Welcome to the Speech Understanding Workshop.   I
have invited the various people working on Speech Understand-
ing in support of the Defense Advanced Research Projects
Agency (ARPA), and other Department of Defense (DOD) and
Government agencies that have research programs underway in
this field.   There are many diverse areas of interest repre-
sented here today; however, all of us are interested either
in furthering research or in using the results of this research
for some operational problem in our various organizations.
The main objective of the meeting, from my point of view, is
that we're reaching a milestone point in speech recognition
where things, not formerly in a realistic sense, are now able
to be demonstrated.   At the same time funding is reaching a very
strained point and although there is no funding center dominat-
ing all the government funding, we believe it is only realistic
to acknowledge the impact that ARPA has on this sort of a pro-
gram.   When the ARPA program terminates in another year, we're
deeply concerned as to what will happen at this point: what
funding we will continue to put in this area and how much

1

funding other people are planning to put in this program. We have always agreed very strongly as to the recognition or we wouldn't be funding it at the current levels in recent years and we very much want to keep technology moving, although it is probably going to be hard for us to fund at the same level as in the past. We see room for a unified discussion and although no firm outcome may come from this, perhaps it will lead to another meeting later on with some of the government people concerned who can talk out future strategies.

The presentations this morning are not intended to be the showing of films or program reviews, etc. However, it is necessary to run through some of the work being done in order to set a base for the discussion. On the ARPA program alone, we could easily take the whole day, or even two or three days, just running through our various programs and different systems. However, we are just trying to get enough information on the table to provide a basis for discussion and, although I'm not as familiar with the non-ARPA programs--there are quite a few of them in recent months--I'd like to ask you all to try to keep your remarks as brief as you possibly can. Touch upon the highlights, the important philosophical issues, without going into too much detail.

Mr. Lee S. Baumann from Science Applications, Inc. (SAI), who has put this workshop together, is also going to provide minutes and will be taking notes. Also, we will be using a tape recorder to assist in the note taking so when talking, please give your name and organization.

Lastly, I'm a little worried about how much time our outside speakers will require. We took a little leeway with the program and did not finalize it until this morning when we could be sure how many people would be here. I think we may have three more people willing to speak than we have time for, so if any presenters can hold their remarks

to ten minutes instead of the fifteen or twenty minutes
allotted, we can make room for everyone. I have asked
Mark Medress, Sperry Univac, to put the ARPA presentation
together into some coherent pattern and to hold everything
down to the minimum time. This has not been easy and it
will require as few questions as possible and even those
few to be as short and as simple as possible. Now I'd like
to ask Bill Woods to give the first presentation about the
speech understanding systems and to outline the other talks
which are to follow.

1.4   BILL WOODS, BOLT, BERANEK AND NEWMAN (BBN), INC.,
        REMARKS

        What I will try to do with the time I have allot-
ted is to give you sort of an overview of what we set out
to do about four years ago with the speech standard program
and what I think we have achieved, where we stand at the
moment, and try to give you a little bit of the flavor as
to what the problem is. Then, at the end, I will tag on
just a little bit of comment about specifics of the deviant
speech understanding system. You'll get considerably more
detail on the SRI/SDC system which will be coming subsequent-
ly.

        There are quite a set of advantages for being able
to use speech as a means of communication between a man
and a machine and this program was launched by the realiza-
tion at various plants that there is just a tremendous
benefit in payoff to get if you could use speech. It's the
most effortless encoding of all the output channels that the
human has available to him, to say things or to communicate
things to other people. It's got a higher data rate than
any other channels you can use, it's the preferred one if
you are going to generate something spontaneously---doesn't
tie up the hands, and one can move around while doing it.

It's just a very nice communication media. Humans, of
course, for centuries have been tieing themselves to the
written record. There are a variety of outward speech of
various types and reading out loud is at least twice as
fast as the record for typing speed and something like four
times as fast as an ordinary skilled typist. It is consider-
ably faster than the average one of us who sits down to a
typewriter and tries to get things out. So, if one could
understand continuous speech, that would clearly be the
preferred mode for an enormous range of situations for a
man who is trying to communicate information into a computer.
Furthermore, there have been some studies that weren't
available at the time the ARPA program was launched that
gives us even a stronger picture of the benefits of using
speech. An experiment was conducted by Oxman and Shupanas
over a wide range of tasks, problem-solving situations,
where one person had to communicate with another person in
their experiment. When they explored the range of communi-
cation channels available to them they found that over a
wide range of tasks - over a wide range of combination of
interactive modes - the problem-solving rate, the speed at
which the task could be done, was enormously improved if
speech was present as part of the communication and not if
it were not. So, in this slide the heighth of the bars
are the average time required to complete the task. The
communication modes are on the left, the communication
range which includes voice - where you can see the person's
face, gestures and everything else - are voice and video,
voice and handwriting, voice and typewriting, voice by
itself, handwriting and video (so you can actually use
gestures to pass them on), typewriting and video, hand-
writing and typewriting, handwriting only and typewriting
only. Clearly there is a very distinct step in the distri-
bution when you drop speech out or when you put speech in.
The conclusions of their study are just enormously strong.

4

The most important single conclusion to be drawn from this research, they say, is clear and unmistakable. There is a sharp dichotomy between modes of communication involving voice and those modes of communication that do not. The dichotomy is characterized by a great deal of consistency within both the voice modes and the hard copy modes. The range of solution time which includes the communication of the voice channel is only 4.4 minutes, that for hard copy modes is 8.7 minutes; there's no overlap between the oral and hard copy modes in terms of solution time. That is, the fastest of the hard copy mode is slower than the slowest voice mode. The data show that, regardless of extra embellishments, communication by typewriter or hand-writing cannot even approach speech in terms of speed or task efficiency. Moreover, these conclusions are consistent and appear to pertain to all kinds of problems and for different tasks assigned to the communicators. Practical implications of these data can be simply stated. The single most important decision in the design of a communication system should center around the inclusion of a voice channel. In the solution of practical and real world problems little else seems to make a demonstrable difference. We didn't really have the benefit of that study when the program was launched but I think there was the intuition on the part of Dr. Roberts that that was, in fact, the case---speech would just be an enormous improvement. As a result of this there was a study to put together a variety of experts in computation, linguistics, language understanding, speech engineering, and phonetics. They were then charged to say, can you build a system with an enormous list of requirements such as a ten-thousand word vocabulary, any number of speakers, real time, noisy input, in three years? And we heard from these people after talking a couple of days. They came back and said, "No, we can't do that at all, but

5

if you make it five years and you make a reasonable set of objectives, there is a good chance that we may be able to do it and it's well worth the risk." And the kinds of restrictions they imposed on the task is a reasonable first step for a five-year program. First is that you must have speech input rather than use any telephone channel. The vocabulary should run a thousand words rather than ten thousand words. Syntax and semantics will be permitted to constrain the things you can say to certain artificial sets, if necessary, in order to have some support from a base line. We will try to deal with multiple speakers but we're not going to try to cope with all the different dialect problems that you get with speakers in different parts of the country.

A variety of issues of that sort resulted in setting up a program that was born recently and we're shooting for understanding speech with that set of characteristics and expectations. Now, in a few minutes, I would like to talk about just what speech understanding is, and the technical problems you have to cope with in order to solve the problem. I'll start with a spectogram, which is probably familiar to most people. The reality that we have to deal with in speech understanding is that there's not enough information in this signal alarm or the spectogram signal or in the other set of parameters throughout the signal, to uniquely determine the phonetic content of the sounds you are talking about. To uniquely determine what the individual sounds are and where the word boundaries are, there's an enormous amount of indeterminacy in the acoustics file themselves. One can see clues in there that tell you I have an unvoiced stop; I can see voiced vowel segments; I can see trauma traction there that can give me constraints. But if a performance breaks down it may be because I have a dipthong where the performance really bends in, or it may be that the performance is bending because of the previous segment which wants to move the articulators with the

mouth closed, or the tongue is out of position, and they
happen to change the performan:e of the preceding vowels.
What one seems to be able to find instead if one tried to
do acoustic transcription from absolute data is that you
can get something like this: There's either an 'l' or 'w'
followed by a front vowel or by 'u', 's', or 'v', - real
hard to tell in order to absolutely say it's an 's' involved.
You get a description that is somewhat partial. In fact,
you get places where you have an optional possible segment
but you're not really sure if that segment is really a
distinctive bend in the signal or whether it's just extra
strong aspiration release from a preceding 't' or some other
phenomena that's going on such as a blip or a variation of
the preceding sound. This translates into taking the future
descriptional source you just had, a list of alternative
possible vowels that might be there; and, from the preced-
ing list, we have something that's either one or a variety
of sounds - possibly vowels, followed by an 's' or 'z'.
It's very clear that one doesn't just easily look at some-
thing like this display and pick out what the sequence of
words are. So, there were a set of experiments done early
in the program that gave us some feeling for the chance of
success. These were experiments with human beings attempting
to read spectograms and do the kind of acoustic transcription
that I just showed you, an example of which is somewhat
vague, and try to uniquely determine the segment or the
possibility of the optional segmentations. And the upshot
of these experiments was, essentially, that trying to do
that acoustical task alone without any syntactic or
sematics aboard is like looking at the signal through a
little window and trying to do just a very objective case
of saying, "does this look like this vowel sound with this
particular consonant". There was about a 25% error rate,
even given that they were able to hedge by saying, "I see

7

one of the following three possible boners". Furthermore, there were several errors in missing - whether a segment was there or not - even though they were able to hedge on that as well. However, in this data there is some trade-off between strategies one can employ. One could try to be very specific and thereby risk incurring a slightly higher error rate because you make more mistakes or, one can be a little bit more cautious in general and specify a larger possible set, thereby running less risk of making an error.

But in a second set of experiments, starting from that, they use a computer retrieval device that would take such a partial phonetics description and come back over to the vocabulary that satisfied it. Then they use their intuition about those words and how those words could be combined in order to induce what's semantically meaningful. In that experiment they found that they were 96% successful in identifying the words. This gave us a pretty good feeling that, at least with human problem solving ability, the information is there in the high level restraints in the language so that we can recover the indeterminacy in the acoustics. An interesting point is that the mystic 4% is almost all confusions between 'a' or 'the' which are acoustically very similar and very difficult to resolve from the kinds of pragmatic information you have in an isolated sentence. There is not much reason for preferring one or the other. So, that's the problem and that's the program that we are essentially shooting for. The experiments which I've just cited, while they relate to human transcriptions and spectrograms, doesn't directly mean that there couldn't be something hidden in the years ahead. The tape-splicing experiments give very good evidence that it's not just a limitation of people reading the spectrograms, but that it's really a limitation

in the acoustical system. As everybody knows, you can say
isolated words in isolation and they're relatively intelli-
gible. So one might therefore conclude that you ought to
be able to do better on acoustics than these data indicate.
However, if you put a sentence, or a word in the middle of
a sentence, people don't say it the same way they say it
in isolation. And if you splice that word out of the middle
of a sentence with a tape splicing experiment, just one
word drawn from the middle of the sentence, the intelligi-
bility doesn't come back up to where you'd expect it to be.
So, you put two or three words consecutively together, and
that kind of context is available to the human perceptual
mind. O. K. The next question, then, is how do we get a
computer to do something that's sort of similar to what
those human spectrogram readings will try. We have to be
able to get the identifying features out of that spectro-
gram that the human perceptual mechanism somehow does when
he looks at it and says, "Oh, I can see there's no voicing
down here" and "his performance is higher than it would be
if it were one vowel sound off so there's got to be some
other vowel sound". How did we get all that into a machine?
The beginning starts with a variety of signal detectors
that can produce the magnitude order slightly better than
the information you have available in the spectrogram.
You can perhaps form the signal itself and this can perform
a tracking which gives you very good approximation of what
you see in a spectrogram in terms of level of performance,
the overall energy curve, frequency, and a spectral deriva-
tive that shows you those places in the signal where the
things are changing most rapidly. There's enormous potential
features one can derive from the signal by signal processing
techniques and each of them has some specific benefit. There
has been a considerable amount of work in this project on
evaluating features that can be extracted from the signal
in determining which features are good, which features are

not good, and which features give you a wider leverage
towards understanding what acoustic signals really are.
Then there is the level of acoustical analysis, which all
systems necessarily do, but at some point it has to get
done from the bottom up to the top down. That determines
what possible segments you see in the other segments that
are not uniquely determined. Here's a typical example of
an ethically derived inventory of all the possible segments
that match well enough. We see that you can have either a
'b' or a 'g', different voice traces of all different poss-
ible vowels, or this entire thing might really be just one
long segment. So the fundamental information that you can
get out of this acoustic circle seems to be something of
this level of determinacy. You may have very strong prefer-
ences, however in some cases, there are only two or three
possible choices. You might have probabilistic expert
patterns but the possibilities of each of these is not just
a statistical decision. You don't have outstanding ability
to say what that could be acoustically, and I can therefore
rely on this being so, never considering the possiblity that
the processing is inconsistent with reality. Think of your
own experience in listening to people, in the speech confu-
sions that they can occasionally make, and you realize that
that indeterminacy goes on in human communication all the
time even though we have very effective devices for coping
with it. If you process this sort of thing and try to
find all the words that you can see and hear you will find
you cannot uniquely determine the words that you can
account for acoustically. This would be true if you could
equally determine what the euphoniums were. Here's an
example from one of our experiments where the spectral
analysis is not as good as some of the others. Somehow
the high-low competency of your system has to be able to
select from all those possibilities that are mutually con-
sistent and go together, that is syntactically correct.

10

It seems very likely that the human perceptual process does this same kind of thing. At the very least you need some grammar that will tell you what possible sequence of words is acceptable. That isn't totally sufficient because you can find perfectly acceptable grammatical strings that are really nonsense sentences. The upshot of this is that a speech understanding system seems to require a lot of different sources of knowledge and has to be somehow integrated to derive what the analysis of a particular sentence really is.

One of the characteristics that results from the problem therefore is that to really work on it effectively requires a special team of people. It requires experts in processing, acoustic phonetics, acoustic analysis, artificial intelligence, and computational linguistics. One of the things that I think is somewhat unique about this assemblage of a group of individuals representing these different areas of expertise is that they must have a common goal and be working together to get something done. A joint program between Systems Development Corporation and a project at Carnegie-Mellon University and a project at BBN are three such teams that have been set up. The main point I want to make is that the speech understanding task has a lot of different dimensions of difficulty. Quality of the speech segments that you get to work on is clearly a determination of how well you are going to do. The size of the vocabulary you're trying to cope with has an effect on how well you can do. The number of different speakers you are going to have to deal with and the dialects you have are important issues, and it is not as easy to qualify signal to noise ratio. I would like to say just a little bit to kind of give you a flavor for this. One could start off with a fairly restrictive kind of language, which dictates that what follows can be composed of only

11

one of these three possibilities depending on which word
is used. The words allow permits such as climb to, maintain
or descend, and what follows are numbers or a few key words
such as altitude or direction.

You can have a language of this sort though its
grammar and word components are specified by essentially
a limited transition diagram that is quite constraining.
It is a fairly understandable technique and easy to imple-
ment. Slightly more ambitious is to allow at various
places in the diagram open classes of words that can become
quite large, such as people's names or names of places or
things. Now instead of your language being able to constrain
the possiblities to one or three or four possible words,
the situation comes up with such a volume to where acoustics
have to choose between maybe a hundred words or two hundred
words and the task becomes more difficult. Further, your
language is not defined by one simple table of transition
but instead you have a set of rules that say there may be a
basic command or word, followed by some variable word, in
turn followed by some constituent which is a class specifier,
followed by maybe, optionally, a word type with the specifi-
cation. Thus, the basic communications system consists of
an operator, followed by several possible variables and a
set of rules to characterize a large class of possible things
you can say. Now it's not even possible to trace out nice
and conveniently through a graph all the possible utterances
you may get. This is an approximation of what you would get
in a high-low programming language or some manageable infor-
mation systems, etc. At the outermost end of the scale there
are attempts to really approximate fluent, natural English.
So, there's a wide range of dimensions of difficulty that
one could tackle. The ARPA group speech projects have been
exploring at various points on the scale. At this point in
the program, I think there are two things that are pretty

clear. There are lots of problems with the lower end of
the difficulties, those close to the category one language
that are clearly going to be solved in a year or so and that
will give a considerable amount of flexibility to some of
those tracts that you really can't get corrected by isolated
word techniques. Second, a great deal of additional research
is required in order to break some of those vicious things
to which I alluded and that is going to take awhile. However,
in those areas, I think, many useful techniques have now
been developed not only for the individual speaker but all
the way down the line in speech understanding. Also there
are isolated techniques trying to deal with speaker varia-
bilities. Some of the specific achievements of ARPA in the
speech understanding project, we think, are really very far-
reaching advances to acoustic phonetic ability, sound seg-
mentations, doing lexical matching from logical rules for
contact articulation, the use of high levels in syntactical
knowledge to compensate for the acoustic indeterminates,
various strategies for incorporating many different speech
patterns and advances in automatic processing techniques.
The many potential applications are well worth the money
spent. I don't think the ARPA program is clearly the final
solution to the problem although significant credit should
be given to the various programs now underway. It's also
clear that there are things not a part of the program that
for many practical reasons have not been dealt with at all;
problems that deal with people from different speaking
heritage, complicated noise and signal ratios, many factors
in multi-speaker models. We have concentrated on those
things that were cost effective to achieve and were most
able to be done.

I've spent most of my time just talking about
speech recognition and I'd like to say first a few things
about the BBN speech understanding system. We do a great

deal of signal analysis to get parameters from many different ones than those that I showed you. We do acoustic analysis to see the acoustic effect on phonetics and we do probability analysis. Also, much effort is devoted to semantic networks and how these might translate into retrieval language.

Editor's note: Mr. Woods' remarks are incomplete due to transcription problems during the last portion of the presentation. A summary of the BBN project is contained in a paper entitled "Motivation and Overview of SPEECHLIS: An Experimental Prototype for Speech Understanding Research", by William W. Woods, published in the IEEE Transactions, on Acoustics, Speech and Signal Processing, Volume ASSP-23, No. 1, February 1975.

A copy of the transparencies used by Mr. Woods are at Attachment 3.

1.5   SDC/SRI PRESENTATION

The next presentation was a description of the work being done by a combined team from Systems Development Corporation (SDC) and Stanford Research Institute (SRI). Mr. H. B. Ritea represented SDC and Dr. Donald E. Walker represented SRI.

Mr. Ritea stated that the goal of the combined team was the development of a speech understanding system capable of engaging a human operator in a conversation about a specific task domain. He stressed that the subject was a joint effort and outlined the specific responsibilities of each party as follows:

SDC	SRI
Signal Processing	Syntax
Acoustic-Phonetics	Semantics
Word & Phrase Pattern-Matching	Programatics
	Discourse Analysis
Prosodic Analysis	
System Hardware & Software	Parsing & System Control

14

Although many systems could be selected, Ritea said, the team finally decided to use a specific data base which contained information on 265 warships of the U.S., USSR, and the U.K. The data was extracted from James Fighting Ships and although not a real world data base the data was, in fact, live and factual and represented near operational type of data. The particular example represented a data management system on the attributes of warships of the three countries selected. He pointed out that the data was all unclassified.

A diagram of the SDC/SRI system is shown in Figure 1.
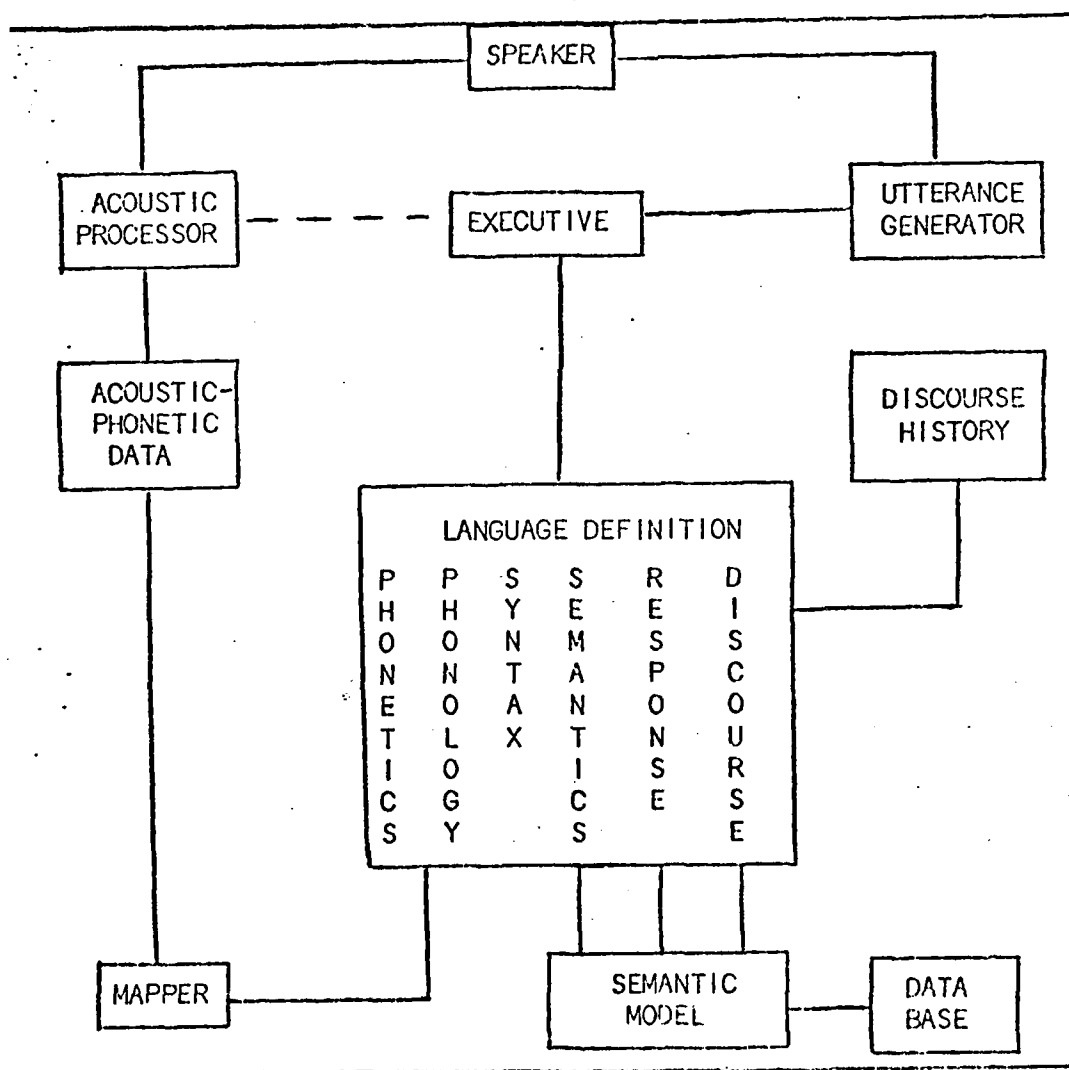
Mr. Ritea then briefly reviewed some of the details of the system to include parametrigation, segmentation and labeling, word and phrase pattern-matching and prosodic analysis.

Dr. Walker briefly described the functions of the executive program and some of the other programs including the Language Definition, semantics, discourse, response, and Utterance Generator.

Ritea concluded the presentation with an outline of the system hardware and software. He noted that several languages were utilized including a new list processing language, CRISP, INTERLISP/370, and the standard DEC operating system, RSX-11M. The three computers involved are the IBM 370/145 for higher-level linguistic processing and word and phrase pattern-matching, the PDP-11/40 for segmentation and labeling, and an SPS-41 for parametrigation.

1.5.1   A copy of the transparencies used by Mr. Ritea and Dr. Walker are at Attachment 4.

1.6   Mr. D. R. Reddy reviewed the Carnegie-Mellon University speech research.

1.6.1   The features of CMU speech research was described as fitting five general areas; general model, multiple systems, automatic knowledge acquisition, performance analysis, and theory.

The general model, Reddy noted, is an attempt to explore many alternative solutions to the speech understanding problem. He explained that CMU has developed three main lines of computer systems; the PDP-10 system for ease of experimentation, the C-MMP using 16 processors, and a PDP-11/40 using microcode for a low cost speech understanding

-16-

system. Allophonic variability, coarticulation, juncture rules and word pronunciation are included in the programs for automatic and semi-automatic knowledge acquisition. Performance analysis is being constructed, Reddy explained, in order to explore various design choices, to attempt to process as close to real time as possible, and to iterate the program design in accordance with results achieved. The CMU speech research is involved in theoretical considerations in language design, complexity analysis and in the study of grammatical inference.

The early experiments using HEARSAY-1 beginning in 1972 were briefly reviewed by Mr. Reddy. He noted that the system concentrated solely on a chess task using a telephone input. Results achieved were 52% sentence accuracy without using semantics and up to 80% sentence accuracy when semantics were added to the program. Running time was estimated as six times real time. A new and far more ambitious program was begun in 1975. This program, entitled HEARSAY-II, has only recently been in operation using a news retrieved task with 15 different sources of knowledge. HEARSAY-II, Reddy stated, is a quantum jump in complexity from the chess task of HEARSAY-I. No results are, as yet, available from this program.

CMU has two speech understanding systems under study using syntax and a lexicon. The first, called DRAGON, was started in 1974 and uses a 194 word vocabulary. Results to date show a 31% sentence accuracy and an 81% word accuracy. However the program runs at 122 times real time. The second program begun in 1975 is called HARPY. This program has achieved 88% sentence accuracy at a run speed of 24 times real time. For the past few weeks HARPY has been run on program language tasks using three speakers with high branching factors. Preliminary results show that on timing

sentences the system gets 80 to 100 percent accuracy on
sentences and 95 to 100 percent on word accuracy. Using
test sentences accuracy falls to 25 to 48 percent for sen-
tences and 75 to 84 percent for words.

Mr. Reddy showed an example of a spectogram using
the sentence "Is there any news about Democrats" from the
HEARSAY-II system. He also implicated that the percentage
of correct identification varies as the branching factor is
increased.

The CMU project has, Reddy concluded, a deep
appreciation of the complexities of the problems involved.
To get high accuracies, he noted, requires careful tuning
of the system with many many runs on training data. Also
although close to real time execution is highly desirable,
results to date show that run times can be very long. At
25 times real time a run can take up to 2 hours, at 250
times real time up to 20 hours can be required. Mr. Reddy
informed the group that systems with many good ideas often
fail because of a few weak links if they are slow. Many
iterations of design choices, he noted, are necessary to
get reliable systems.

Mr. Reddy concluded his remarks by showing a
composite of the results achieved with various systems up
to this time. The results of effective vocabularies used
by various systems is shown in Figure 2.

1.6.2   A copy of the transparencies used by Mr. Reddy
are at Attachment 5.

This concluded the review of the three Systems.

18

| | LANGUAGE | | | CONFUSABILITY | |
TASK	SIZE of VOC.	ENTROPY	EQV. BRANCHING FACTOR	ENTROPY	EQV. BRANCHING FACTOR
DIGITS	10	3.32	10	0.24	1.18
ALPHABET	26	4.70	26	2.43	5.39
ALPHA-DIGIT	36	5.17	36	2.29	4.89
CHESS	31	2.87	7.30	1.73	3.32
LINCOLN	237	2.84	7.18		
EXTENDED	411	3.36	12.61		
IBM	250	2.872	7.32		
PROG. LANG. (No Syntax)	37	5.21	37.00	1.92	3.78

Figure 2.   EFFECTIVE VOCABULARIES USED
BY VARIOUS SYSTEMS

1.7   The next part of the program focused on the work of specialists as contracted to the systems work. Mr. Frank Cooper from Haskins Laboratories acted as moderator for the specialists representatives.

1.7.1   Mr. Cooper noted that the specialists program consists of four smaller undertakings. In the original planning, he stated, it was decided to proceed with feasibility tests connected with speech systems, building on what was then known about acoustic phonetics. Prosodics, stressed intonation phonology and looking at all the kinds of information that it was hoped to extract directly from the acoustic signal. There was, however, a recognition that more would be needed in these areas. The system builders would be expected to do research on the parts that they needed for their own systems. But, he opined, additional effort was needed for two reasons; first to supplement and assist the systems builders but also to lay foundations for sound generations that aimed at something beyond feasibility testing. For this reason, the specialist contractors have been working mainly on how to milk more information out of the acoustic signal, with an eye to achieve some assistance but also working with the systems builders to put these findings into use.

Following his opening remarks Mr. Cooper introduced the research being conducted at Haskins Laboratories. He noted that Haskins work was motivated by some work being done on speech reproduction. Looking at how speech is generated, he said, you find whole sentences or phrases, and these necessarily break into words. Everybody is aware of this. The next unit down the line is essentially the sole, and these are the characteristics of all the English language. It is the sole, the base unit upon which the whole phonetic structure depends. When reduced in detail, it seems to be

the most tightly restructured unit that is generated in connected speech. The accommodation between adjacent signals is generally called the coarticulation task. The sole is also the carrier of stress. It's the signal where the pitch rises and changes or which is strengthened out or is found in a much reduced form and yet retains its basic characteristics even if it doesn't look the same. Our approach, he noted, was to parallel the general efforts ongoing but to go down the line to find phonetic or phonemic units that looked alike. First, we made the main cuts where syllable boundaries occur and then to see what could be done about individual syllables because the main thing about a syllable is that it has a very limited structure. It has a vowel or something like it in the middle, possibly glides around the vowel and may have a closure type consonant. The point is that if you can find out anything about a syllable you immediately know some of the other possibilities that exist. You can go from the outside towards the middle or start in the middle and work back to the outside. At any point along the path you have much reduced the possibility of things to look for and might, indeed, consider a pattern-matching operation on a small subject of syllables. This view is quite consistent with other peoples' ideas. We are simply trying to reduce the task by taking account of what we know about the acquired syllables in English.

Following these remarks, Mr. Cooper introduced Paul Mermelstein to conclude the Haskins presentation.

1.7.2   Dr. Mermelstein presented the following description of the Haskins Laboratory Projects:

Our general program is oriented toward natural speech and how these cues can be applied to speech recognition and speech understanding. We are working now with

21

two lines. In one line we've been considering some algorithms
for acoustic feature extraction of the speech signal. One
portion of that is to support the system codes in general
and systems developers such as the SRI system in particular
as far as aiding them with particular feature extraction
algorithms. We've done a certain amount of work on human
visual analysis of the speech signal. Here our philosophy
is that by transferring the analysis from the auditory mode
to the visual mode many of the processes that go on un-
consciously when the human perceives a speech signal can
be explicitly formulated. For example, acts of context,
rate of speaking, and stress, in changing from a reference
form as it may exist in a lexicon of words or continuous
speech form a major study effort. And finally, the work
of the organization of feature extraction process is primarily
how to put together, how to combine, the information from
several features extraction outlets and, in particular,
based on human visual analysis, to find out if hierarchic
rather than parallel systems of features extraction may be
better. This is based on systems features that have initially
been analyzed, and then selecting what other features to
look for rather than looking for all features in general.
As a specific, our machine algorithms for feature extraction
are done by cementing the continuous signal into what I
call 'flobic' units. These don't quite correspond to
syllables because the boundaries may not be precisely where
one would put them on the basis of a logical continuum.
One study that's been done with the Systems Development
Corporation vocabulary shows that the syllables correspond
roughly to 1 1/2 words for any unique syllable that we could
have found. Therefore knowing what syllable there is in the
acoustic stream we know quite a bit about the possible words
present. It seems that syllable segmentation is much easier
than segmentation into individual words. There are several

other things that people are doing not necessarily as part
of the ARPA work but using information that we have devolved
through that work. Some work has been done on improved
analyses tools. One of these is the additional patter play-
back where we have concentrated on aiding the human experiment
to increase his context retrieval.   This gives a facility
to modify the spectral information present in the signal and
then getting record feedback as to the full importance of the
specific cues noticed in the signal. As you may know our
phonetic events are accompanied by a number of diverse acous-
tic cues and it's not always easy to tell which of these
carries the significant information. Thus by selectively
adjusting these cues and presenting them to the experimenter
he can then study which of these is the most robust and
which of these he may want to build his algorithms on. Other
activities that I want to mention are primarily the speech
perception-speech deduction studies sponsored by the NIH,
speech synthesis work sponsored by the VA, and the support
for play-back pattern development sponsored by the NSF.
We have attempted to combine our speech understanding from
the combined results of the information that has been
brought forward through all of these studies.

1.7.3   A copy of the transparancy used by Dr. Mermelstein
is at Attachment 6.

1.8   Mr. Cooper then introduced Dr. Wayne Lea who pre-
sented the Sperry Univac part of the specialist program.

1.8.1   Dr. Lea's remarks are as follows:
         The Sperry Univac school for ARPA has basically
to do with analysis of prosodic features to determine how
they might be used in a speech understanding system. We,
essentially, have a twofold goal. One is to find analyses
tools and test them out in a system. This involves taking

such things as fundamental frequency, energy contours and durations of segments and turning these into abstract pro- sodic information such as phrase boundaries, stress patterns and intonation rhythm. Then that kind of abstract prosodic information is turned into aids to word-matching, parsing and segmantic analysis in a system. What we have accom- plished in this area, basically, is to build the tools. We're just now at the stage of starting to see how these tools could be applied to an actual speech understanding system. In addition, we're doing research, experimental research, that will help us build better tools. This in- volves various experiments on prosodic structuring and so I'll describe both our research and our development of tools. Let me say a little bit about the importance of prosodic analysis; why we see it as relevant to the speech understand- ing system. There are various ways in which protheses can reduce computations. One example is to only do detailed spectral analysis in the stress syllables which we know are essentially islands of reliability in the connected speech. A second point is that we can determine sentence type from prosodic structures. For example, we can estab- lish that a 'yes' or 'no' question has a rising intonation at the end. Then we know something about the possibility of one sentence type versus another, independent of what the word content of that utterance may be. We also have the ability to disambiguate sentence structures breaking them up into phrases, and establishing that one particular reading of a sentence was intended instead of another. We hope to be able to be involved in using pauses to break a connected discourse into manageable size units, sentences and clauses that can be functional in a speech understanding system. I've listed here a number of ways in which stresses provide information for speech understanding. They are, as mentioned, islands of phonetic reliability, and I'll

mention some experiments to demonstrate this in a minute. We also find that stresses provide us with the most important words. If a word is important, then it's stressed. We find that there's a closer connection between the phonetics, acoustic phonetics and the underlying phonemic structure in stress syllables than there are in reduced syllables. We can subset the lexicon, so to speak, by saying at a particular point in an utterance that only those words that have stress in the right positions can be hypothesized at this point. There are other ways in which stress is important. One is the condition on many of the phonological rules that have been developed in the ARPA program. One particular way in which to use a condition is in the area of rhythm and rate of speech. Now we know that when you speak faster there's more slurring of soles than if we're speaking slowly. Therefore, if we can establish rhythm and rate then we'll have something to guide us as to which phonological and acoustics phonetic rules might be applied in the system. We find that the time interval between stresses is the best cue to the rhythm and rate. We also can establish something about syntactic categories from stresses. For example, the contrast between a compound noun and a nuclear noun phrase in the syntactic structure allows us to find something out about subordination. When a phrase is subordinate to another one its stresses are at a lower level than those in the superordinate phrase. Shown here are some highlights of some of the experiments and these are only a few of the experiments dealing with stress and boundary marking in prosodic patterns. We find that by looking at a number of the techniques for machine classification of speech into phones, the segmenting of speech into phones and putting price labels on those phones, that these techniques, regardless of where they're done around the country, are working better

in stress syllables than they are in the other portions of
speech. This, again, is a just a way of reiterating what
our intuition says, mainly that stress is the place to find
reliable information. Another point is that listeners can
consistently find stress, or which are the stress syllables
in connected speech. We find only about five percent con-
fusions from time to time or from listener to listener in
this area. This says we have a standard for establishing
where the stresses are in speech. Now the question is, can
we find acoustic correlates of stress? And the answer is,
"yes", we find that previous experiments have demonstrated
that stress syllables are accompanied by rising fundamental
frequency, high energy and long durations in surviving
nuclei. From that we've developed a computer program which
I'll mention in a minute. In addition, phrase boundaries
are detectable from prosodic structures in one of several
ways. Independently you could use the fall/rise values in
fundamented frequency contours to find boundaries between
major syntactic units. You could also use long-time inter-
vals between stresses. It turns out that the long time
interval between the onset of two stress syllables, or stress
vowels, is longer when there is a syntatic boundary between
those stresses. In addition, there is lengthening of the
vowels and consonants at the phrase final positions in the
utterances. So, here we have several ways of finding phrase
boundaries from the acoustics independent of the exact words
that are in that structure. From this kind of research we
have been able to come up with a few tools for prosodic
analysis that can be incorporated in speech understanding
systems. These have been provided to the ARPA system build-
ers and in particular all four of these are being explored
in the Bolt, Banareck & Newman system. We have a fundamental
frequency tracking algorithm similar to the one at SDC and
also similar to one at Bolt, Banareck & Newman. We have a

method of detecting phrase boundaries from fundamental
frequency contours. This works about 90% of the time in
finding the phrase boundaries that we would expect to be
there based on independent syntactic analysis. We have a
procedure for syllabicating speech much as Paul Mermelstein
has described. It is a little simpler, and gets about 91%
of the syllables in connected speech. We have a program
for finding out which are the stressed syllables and this
works about 89% of the time. These programs are now avail-
able and they are tools. The question really is, how do
we use them in a speech understanding system? We are cur-
rently trying to relate them to a parsing process in an
automated transition network effort at Bolt, Banareck &
Newman. We're going to try to work on the guiding of a
parser for doing the most efficient analysis and see if the
prosodies, in fact, can help in that area. It is worth-
while noting, particularly in the light of the variety of
interest that is represented here today, how the speech
understanding work has contributed back to other work. For
example, Sperry-UNIVAC's interest in speech recognition is
in more limited form, as I think some of us here are today,
and we've developed several systems that are involved in
restricted speech recognition that are using a lot of the
tools that are coming now to the ARPA program. In particular,
we have a prosodically guided word-spotting strategy that
is finding key information carrying words in connected speech
and that strategy is using prosodic guidelines that we de-
veloped for the ARPA program. We're using linear predictive
analysis pretty much as it's being used by other ARPA
contractors based on their ideas. We're using phonetic
classification schemes that are based on all the work going
on in the ARPA program. We have the segment lattice notion
that you heard Bill Woods of BBN talk about. We have a
word-matching and scoring procedure which is very similar

and somewhat based on the Lincoln Laboratory ARPA system. So, also, in several ways is that system for word spotting using the output out of the ARPA program. We also have two restricted speech recognition systems for recognizing isolated words and connecting word sequences. They, too, are trying to use the linear-predictive analysis and the phonetic analysis techniques out of the ARPA program. We're dealing with co-articulation rules and other aspects of word-matching that are coming out of our own program. So we have a two-way street here in which information is being provided to ARPA by the research on prosodies but we find that the ARPA program is also providing good tools for immediate work in speech recognition. Thank You.

1.8.2   Viewgraphs shown by Dr. Lea for Sperry Univac are at Attachment 7.

1.9   Mr. Medress explained that the next scheduled speaker was Michael O'Malley from the University of California at Berkeley; however, Mr.O'Malley was not able to be present at the meeting. Medress asked the group to note that O'Malley is a specialist contractor on this program working on prosodics emulations to syntax and other aspects of speech understanding systems.

Mr. Medress then introduced the last speaker in the ARPA group, June Shoup from the Speech Communications Research Laboratories.

1.9.1   Dr. Shoup's remarks are as follows:

I will very quickly give the objectives of the SCR project for the ARPA work. First of all we had endeavored to accumulate a large natural language data base and then have these transcribed orthographically, ARPAbetically (I'm sure you're familiar with what ARPA code is, pseudo-phonemic code) and also phonetically. Second, to develop

28

computer programs for analyzing the above three levels of transcription. Third, editing our computerized natural language in the SCRI dictionary which contains orthographic communiques and gross grammatical codes. Four, compile phonological rules to obtain natural language variability from the dictionary base form entries by analysis of the natural language data at the phonemic and phonetic levels. And, sixth, support for the ARPA speech understanding system builders. Now, very briefly, the accomplishment toward these six areas.

First of all on the data base, we did obtain over 200 twenty minute tape recordings and we reproduced them at all levels of signal to noise. We have some in anechoic chambers and then we have some in households where there are children in the background. We have started with those with better signals to noise ratio for transcribing and eventually hope to get to the worse ones. We've taken over 30 of these and transcribed them orthographically and ARPAbetically. We have only done a few phonetically. One of the things that we have developed is a study on how accurate you can get phonetic transcriptions which are intra-speaker and inter-speaker. This study shows what the techniques are and how reliable the phonetic transcription work is. Part of our work has resulted in very accurate results and we have actual-ly transcribed and put them in the computer. On the computer programs themselves we have now developed a complete set of programs to categorize, reference and analyze the various levels of transcription including a cross reference of data. We can ask most any question regarding what words have certain phonetic features, are spelled orthographically, or what grammatical culture are associated with certain phonemes, and we can go back and forth in anyway throughout the data. Regarding the dictionary, we have completed the editing and we have written computer programs for updating and

maintaining the dictionary, which now contains hundreds of thousands of entries. We will always find missing words and will have to add them. We, therefore, have our programs for updating and maintaining the dictionary. In the area of phonological rules, we have completed obtaining all the rules that were in the published literature and have started testing these against the natural language data base. Obviously, these were not complete nor correct, so we have continuously modified and added new ones as we have analyzed the data base. For the analysis of the data, we have just initiated statistical studies on consonant clusters, vowel clusters, the phonemic substitutions, deletions and additions, the phonemic frequency in various positions, and the phonemic-phonetic comparisons. In systems support, in response to requests by the three systems builders, we have provided ARPA transcriptions based on dictionary entries for their individual dictionaries. Also, we have done rule testing as it applies to their data and have written a common task report. This report is available as a technical report if anyone is interested. The task was to ascertain whether it was reasonable that the three system builders have a common task to perform during this last year or whether they should just remain with their present tasks. The conclusion is that it is not feasible to have a common task.

Thank you.

1.9.2   Transparencies used by Dr. Shoup are at Attachment 8.

1.9.3   Mr. Medress announced that this completed the review of the program from ARPA's prospective. He noted that several of the speakers had brought along copies of papers or background materials. These, he stated, would

be available on the table at the lunch break. Also, Medress concluded, additional materials may be secured by writing directly to the appropriate speaker. He then turned the program over to Major Carlstrom.

1.10   Major Carlstrom surveyed the non-ARPA speakers who had prepared material. He then introduced Commander Wherry from the Naval Air Development Center, Warminister, Pennsylvania.

1.10.1   Commander Wherry presented the following remarks:

The voice recognition and synthesis program at the Naval Air Development Center started in complete ignorance of what was going on in ARPA. We have developed a system, probably in isolation but, fortunately, it works. It does some things which we think are fairly neat and, by way of background, let me tell you how we started into this program. The Navy had been funding some research on voice recognition with Scope Electronics Company, and they had showed some fairly good voice recognition accuracy rates in quiet rooms. So at Pax River one of the pilots went up with a tape recorder and reported in while he was flying, pulling 'G's, and doing maneuvers, etc. We brought the tape recordings down and played them into the device and the voice recognition accuracy rate went to pieces. And so the Navy wanted to know why. They put out a call to various laboratories and at the Naval Air Development Center we happened to have a 50 foot centrifuge in which we can control pressure levels and wear oxygen masks at high vibration levels. So, we won the research grant, if you will, to try to isolate what it was about the voice quality that must have been changing in the air environment that resulted in this horrible recognition rate. So we did the voice quality studies which I'll

get into. Then I'm going to talk about fabrication of the VRAS Facility which we have accomplished at NADC and in the development of message understanding software which has all been geared to work a demonstration of VRAS applicability in airborne laboratory systems. The voice quality studies which were done back in Fiscal 72 covered several major elements that we thought would be present in air-warning applications. They included cockpit temperature, vibration levels, G levels, whether or not the man wore a mask or did not wear a mask, the mission duration, the number of words that he had to say and the noise level. Noise level had an effect, but it was not something that was getting into the microphone for we played different levels of noise into his ear. Now, what happens when you do that is the level of the voice goes up when the level of noise coming into the ears is high and voice quality does, in fact, change. This was one of the things that the investigators didn't get on the transferrency but we did look at it. Our major findings were that voice level tends to degrade if you're wearing an oxygen mask after about half an hour. Now we had some slight positive pressure breathing and one of the things that happened is that it does tend to exhaust the voice. It turns out that voice recognition is better during the first half an hour by virtue of wearing a mask, because it cuts out some of the ambient noise and probably results in better formed words to the machine. A second thing was that under high vibration levels, .35 I believe we used, or .39, we found some degradation in accuracy rates. Under about 2 G's we had about a 6% to 7% degradation in accuracy and at 4 G level we had about a 12% degradation in accuracy. Now we feel that a lot of this reduced accuracy is in fact attributable to the mask that the man wears. Under vibration and under high G's just the mass of the mask itself begins to pull down and you get some nasal qualities and the

voice does change, so the machine has some difficulty in recognizing it. These were the major findings. We decided on the basis of that, that there were, in fact, some aircraft in the Navy's inventory that we could put voice on, probably immediately, things like aircraft which don't undergo high levels of vibration, where you do not have to wear masks, and which do not undergo high G's. So we started building up a facility. Our facility consists of the voice analysis portion of the system with a Scope DCS playing into a raytheon 704 computer.

We now have 32 K of core and we have a disc drive where we store off-line different speakers vocabulary sets and we use a Votracks as our voice synthesis unit. We also have a Burroughs self-scan which gives visual feedback to the operator in terms of what word he's saying or what the word means, or is one of the ways that he can ask for information to be displayed to him. In addition to that, we are setting up an interface into a simulator, an aircraft simulator at NADC. Looking more schematically at what VRAS consists of, we see it having to consist of some portion of voice analysis which feeds into a statement-understanding box which, once the statement is understood, goes into a message-handling box. The message, which is either to gather information or change the state of information somewhere in the system, then passes over to the system computer; does what it has to do, comes back, and a response is generated which can either be put out by voice or visually. So this is basically what we consider to be a voice recognition and synthesis system. Let's look inside at a couple of those boxes. In terms of the voice analysis box, this is the same scheme that is used by the Scope Company and, I think, by some other companies, but not by all companies. And, essentially, what is happening here is that from the microphone it comes into 16 band-pass filters and is digitally

chopped every 60th of a second. So, we've got a sampling
and a word boundary sector. Now there would be, of course,
other ways of doing this. Many have been mentioned here
such as looking for syllables or somehow breaking the state-
ment up into words or syllables or whatever you have. But,
there are amplitudes that have gone through these band-pass
filters which we call the long vector, and are capable of
playing one second duration words. Now, of course, in our
system we're working in isolated words. We're not particu-
larly worried about the different speaker problem because
we know which speaker is sitting at which microphone in the
aircraft that day. So we haven't broached those kinds of
problems, speaker identification problems or anything like
that. When the word boundary detector says, "Hey, I've got
a word", then we go through a word compressing algorithm
and come out with one hundred twenty Bits which are then
passed on to the statement understanding block of the pro-
gram. And I want to say that, really, the majority of the
work that we've been doing has been put into this box.
Essentially, the spoken word comes in and we have a word
correlator subroutine which is comparing the incoming word
with a list of words that the machine thinks he could be
saying now and we come out with the highest correlation. If
we have one that is high enough, we put it into a block of
a word that's been heard. We then go to a translator section
which goes to look up the meaning of th-t word in word-mean-
ing dictionary and put it into what we call the matrix of
meaning. We'll discuss this in a little more detail in just
a minute. After the translation phase of it we go into an
entity eliminator. Now what we're doing here is saying, "by
the meaning of all the words that I've heard so far, what
do I know that I won't hear from now on", and so it does,
in fact, eliminate future words that we would have to compare
against. Having done the interview elimination we come down

to a statement testor to see if we've got sufficient information that that is, in fact, the end of the statement. If it is, then the message is complete and it is passed to the message handling box. Otherwise, we come into the word selector which says, "knowing what we know about things that have been eliminated from consideration which of these possible words in the total dictionary could he now be saying". Then we form a list of words and essentially keep running around in this loop until we've got a statement. If, at any time we don't, we fail to understand a given word, then the machine starts saying, "Say again all after.........". In the terms of the word types that we've identified, we actually have 19 columns of information in what we call the universal statement. Some of those word types are used only in the responses like preambles, things like "Sorry, I can't do that", or "Accomplished, I have done that", or "Affirmative, the answer to that question -", or "Negative", etc. There are other action kinds of words. We want information changed, or we want information gathered, words like report mean "gather some information and send it to the talker unit". We have identified what we call post-verbs using words like, be, that, this, any, another, other. We call them post-verbs because typically they follow the verb, and they do indicate the version of the thing that he's talking about. Now, when we talk about things, in at least aircraft systems, we're usually talking about objects that you can actually point to. Therefore, a thing by our definition is a system, or a sub-system of a system, or a component of a subsystem of a system, or a version of a component of a subsystem of a system. Thus, essentially what we mean by a "thing" is something that can be pointed to. Now most airborne operators may deal with a variety of systems. Usually he will deal with more subsystems in his routine activities than total systems. For example, I mentioned the characteristics, that when we need something

like "location" that we would be talking about the specific
characteristics of the X or Y location, or altitude, or range
and bearing. All these deal with location as an attribute
of something. We may be dealing with motion as a dimension,
in which heading and speed would be a characteristic. Essen-
tially we are saying that all the information which the
operators pass back and forth to the computer deal with
attributes about things. Then we have logical operators,
phrases like: is, not, equal to, or less than. Then we
also have value scales or destinations; the values of scales
are obvious, the destination where I want the information
sent to as the talker, the scanner unit or the printer unit.
Now, in terms of what we call this universal statement, what
we really are saying is that all commands or requests
essentially began with some word that says either, I want
information gathered, I want it compared, or I want it
changed. Now we will allow many synonyms to mean that,
but essentially statements start with things like that if
it is a command or an action. There are some conditional
statements like, if something is true - then report altitude,
if altitude is less than 2000 feet - then report altitude.
In terms of the responses, our machine is intelligent enough
to know what it can do and what it cannot do. So if you say
to report such and such a piece of information, it may know
what you are talking about but it may not have sensors
tacked into that piece of information. In this case it
says, "Sorry, only so and so, or X is only known by so and so".
It essentially tells you where to go to get that piece of
information or to get that piece of information changed. If
you ask it a question like, "Is altitude less than 2000 feet?"
it always responds with the positive information that it
knows about. Finally, communications sometimes do not deal
with the information that operators want to exchange with
machines but they deal with the communications themselves.

36

Therefore, we need words like "repeat" which means - say that last statement again to me. If the machine does not fully understand what it is that you have said and wants to confirm it, it says "understand......", or it says things like "say again all after". So, really, we split our world into one universal statement and say that all statements must have all of these components. Now how do we get all these components in? It is, in part, through the dictionary of meaning. We can take a word like "enable" and tell the machine in its dictionary of meaning that "enable" means change selected status to on. Now we don't have to say all of the words, change status selected status to on, because that is what "enable" means. In this way we can cut down the number of words that appear in the statement. We have not outlawed it from him saying "change selected status to on", but merely that there is another way we can approach that problem. When a word is recognized as "enable", and the dictionary of meaning is looked at, then the appropriate words are put into the appropriate columns in the matrix of meaning. To us "radar" means "the Helo radar" in some particular application that we are doing. We put one other feature into the VRAS system. This is that for this particular operator, when he talks about radar, he is usually talking about his radar display. He is not talking about enabling the radar equipment, he is talking about enabling the radar display, because he is a tactical officer and he is looking at that display. So, if he doesn't tell you anything more, then we can use the fault words which appear in the dictionary of meaning. That is, essentially, what happens except one other feature that we have added is what we call a truth matrix approach. Each time a word is entered in like, "Helo" or "radar" or "display", we eliminate the need to look at so many different words. This approach is one that you may

be more familiar with in puzzle magazines where they say,
"Jack and John and Joe are married to Mary and Alice and
so-and-so.." You have all seen those kinds of things.
Horrible! Well, a truth matrix approach says that when
people communicate with one another they don't always tell
you specifically what it is that they are talking about.
But, they give you sufficient information and you ought to
be able to figure it out. So, essentially, we put the truth
matrix approach into the VRAS system so that if we see a
system word being entered into the matrix of meaning, we go
into our things table, which lists all things that this guy
can talk about, and we eliminate everything that doesn't
have that system connected to it. If we see a subsystem, we
do the same thing. That is, we eliminate all things that
don't have that subsystem. So, once we have eliminated a
given thing, we come over and eliminate it from our entity
table which is the intersection of things and attributes.
Then we look down into the attribute table and say, "Can we
eliminate any of these attributes because they no longer
exist in the entity table?" The example that I used in
terms of enable was something where we actually learned
the attribute before we learned the thing. We can also
go in the reverse direction. We can say that he is talking
to us about status, therefore eliminate all attributes that
are not status. Thus, come over to the entity table and
eliminate all entities that do not have statuses and then
start looking for those system or subsystem words that
do have statuses. Essentially that is the VRAS system. We
are always looking to try and reduce the message confusion
and I will say that the probability of getting the message
understood is the joint product of the probability of recog-
nizing each word as it is said in the message. We, by
introducing rich meaning words into the dictionary, can
reduce the number of words in a statement. But this increases

the number of words in that list. So, there are always trade-offs that have to be looked at. Essentially this is the kind of a program that we have developed at NADC and it does work. The interface in the simulator is not yet complete. A couple of other features is that VRAS also remembers the last statement that you said. For example, if you have just asked for speed, you may now ask for a report on heading, and it will very rapidly home in on that. You do not have to put all of the words in the message. It also remembers the last version of everything you have talked about. If you have once identified the last thing you talked about as Buoy, then you discuss something else for awhile, then you want to go back and talk about that Buoy, all you have to do is say "the Buoy" and it knows which one you are talking about rather than having to say "Buoy two four" or "Buoy 24". And one final kind of thing it does is to recognize that not all things should be changed by a computer because it might misrecognize what you said. So one should very well plan to confirm certain things. It knows which entities it must confirm before it can execute these statements. And, in regard to that, we have one other slight feature. A man can make a statement which does not end with the word over. In this case he can time out on it, that is, not say any more words, and after awhile the computer says, "I bet he is done". The statement tester then says, "Yes, I have enough words to make a statement out of it". You can also say "confirm" in which case the computer will come back with "understand" and repeat the words that it thinks it has heard. Then, if the man says "affirmative", it goes ahead and does the statement. The other thing is that sometimes you don't want to have to wait for a very long number of words in a statement to get something done. So we incorporated the feature of standby. The man can say "standby" and the computer says "ready", and then you can make a statement.

Then it says, "understand......." In this case you are doing
a confirmation early. Thus, you would say, "affirmative" and
it says "waiting", and then it takes that statement and puts
it into the buffer somewhere. It knows what you have said, that
that you confirmed it, and that it understood what you said
properly. Now, at any point you can say "execute", and it
does the entire statement. That is the VRAS system.

Thank you.

1.10.2   Copies of the transparencies used by Commander Wherry
are not available.

1.11   Major Carlstrom then introduced Dr. Bruno Beek from
the Rome Air Development Center at Rome, New York.

1.11.1   Dr. Beek's comments are as follows:

I will try to keep my remarks quite short. Speech
understanding is just sort of a subset of speech recognition
programs that we are interested in. We have been doing quite
a bit of work on speaker verification and things that you will
hear a lot more about today. What I won't talk about in great
detail is automatic message monitoring. We will be talking
about surveillance communications channels. There is quite a
need in the services for that. The other thing that we will
be talking about today is voice input. I put together a
paper with Dave Hodge and Ed Neubeurg where we talk about
military applications, possible applications, and coordination.
This is the table from that paper. If you are interested in
it, write me and I will send it to you. Essentially what it
is saying is that we are interested in many things as speaker
verification and identification. We want to recognize spoken
and digit codes. We want to use voice as a method of access

control and we are very interested in the command and control functions for voice input. In the services, we have to work with requirements and, for example, these are a set of TAC and SAC requirements and Air Force Security Service requirements. The first two deal with verification systems. The third one deals with airborne skint systems that is tied in with the process of message monitoring. Also, we have some work to do on a requirement that was put out by the Air Force Flight Dynamics Lab. We are interested in expressing human speech quite similar to what Commander Wherry was talking about but, specifically dealing with what happens to the speech mechanism under High G forces. We also do a great deal of coordination and here is a list of other people who are quite interested in automatic speech processing, not necessarily in speech understanding, but in other programs. These are a few other technical committees working in the area. This comes directly from the report I mentioned. What I would like to say here is that we, in the services, are already building speech recognition equipment. Some of them are being processed for field use now. There is quite a bit of work in narrow band speech communication systems and in speech compression which I think most of you know about. We are also doing work in automatic speaker verification. I will show you a model of that. That is undergoing operational testing now. There is quite a bit of work going on in training systems and limited speech understanding systems. It is mainly isolated work that they are doing. SRI has an isolated word recognition system that they are using in a training system mode. There is also some work in the Army as well as by others. Helium speech-- quite a bit of work on that. On-line cartographic processing--the cartographers for the Defense Mapping Agency have a real problem. They have to get data into the machines and using the manual mode or keyboard entry just doesn't fit

as you have to do it much faster. Another System that is being built is the word recognition for military tactical data systems put together by the Army. I don't know if Mike Simpson is here but if he is not, he wrote a very good paper that will be published at the Conference on Artificial Intelligence, later changed to Conference on Interactive Cybernetics - because 'intelligence' is a bad word right now. Also, I think Jim Glenn is here from Scope who is actually putting the system together; so, if you are interested in that you can talk to them. Voice recognition and synthesis for aircraft cockpit, I think Commander Wherry talked quite a bit about that. We are going to start looking, in combination with Flight Dynamics Lab, at the problem of what happens to speech under high G forces. We are not trying to put a box into the system but actually trying to understand a little bit more about the speech mechanism. Here is a tutorial; a view of things that we do up at Rome. We work at message monitoring. We are working in voice control systems and we are doing some work on cartography. We have built a speaker verification system that is now undergoing operational tests by Electronic Systems Division. Now, message monitoring also encompasses doing something about noisy, corrupted speech signals. We've done quite a bit of work in that area and there is a great deal of interest. These are the three paths that we are looking at. One is - can we recognize when a voice modulated signal has occurred to enhance the speech and the signal to noise ratio, and can we use this system as a preprocessor for all automatic systems? In general, we found that most automatic systems are highly corrupted by noise and distortion over communication channels. We notice that there is quite a bit of controversy and in the tests that we've run we found we have many problems in that area. As to message monitoring, it is listed here as key word detection but you can substitute language detection, or

speaker detection. What you have is a set of intercept receivers, and you would like to do some sort of automatic monitoring of these receivers. By switching the matrix you can actually send it to the linguist who will be interested in that particular signal. Shown here is the automatic speaker verification system. In the center, we have the system that was delivered, and in the periphery we have a scenario of how the thing works. It has been working for over a year at Texas Instruments and it works as follows: The individual comes into a booth, he sees the microphone and a badge reader, and he inserts his badge into the badge reader which identifies him to the machine. Then, the machine talks to him and asks him to say a few phrases, which he does. If he is accepted, he is allowed into the room and if not, he is locked into the area. This system, as I mentioned, has been working for over a year. It works on a data base of over 170 speakers. We have learned a great deal about this type of application. For one thing, you have to be concerned with more than just the technology it-self--you have to worry a lot about the system's human acceptability. We had quite a bit of problem with the system at first which had nothing to do with technology but with system acceptability. For example, when we had a power failure all the doors were open. In addition, we found that we learned a great deal about test and evaluation of these systems. You have to do more than just one speaker or 10 speakers, you have to talk in terms of hundreds of speakers. I now want to talk a little bit about performance. The base and installation security system at ESD did actually come up with some performance specifications which said that the true speaker could only be rejected 1% of the time, and imposters could be accepted only 2% of the time. Using the data with one phrase you can see right here that we are not able to make it on a one-phrase system. But, when we use

43

two or more phrases we were actually able to exceed their specifications. Right now these are the major features of the system. It actually uses 16 different words which can be put together in a random way so that it makes up 32 different sentences because people are worried about what happens when an individual brings in a tape recorder. One way of defeating it is by having random words coming up. It uses voice prompting, as I mentioned, and, in addition, uses a sequential decision strategy which says that if you do very well on the first phrase it allows you in immediately. Over 75% of the people are admitted on the first phrase, and on the second phrase over 96% of the people are admitted. Right now the error rate is about three-tenths of one percent as compared to the one percent required and the imposter acceptance is about one percent. Another problem area is this one. The map-makers have to extract data and you can see it is longhand printed, and has other types of symbols on it. There is overlapping and optical character recognition just cannot handle this job. So what can we do about that? There is one system that is in operation right now and it is a threshold technology system. There are others on the market but we find this one quite suitable for this application. We found doing independent testing using our tapes that the 10 digits plus five control words has an error rate of about one percent. However, what we are trying to determine now is, again, user acceptability. So, we are going to have cartographers run the system, and they are going to compare the data that they had with hard manipulation so that we can actually get performance characteristics of different systems. Here is a quick run-down of what this system needs. The Defense Mapping Agency system has test digits, it works on-line, and has all the good things that these voice recognition systems can do. One thing I should say something about is isolated word

recognition systems. They are sort of being down-played
today but, I think, they can solve some real important
problems. These systems work with high noise background;
they also have rejection capabilities so that if you say
other words these words are rejected. This will solve a
lot of these type of problems including day-to-day variations.
At first, when these isolated word recognitions systems were
built, if you trained at the same time you would use the
system and they worked quite well. However, as time went on
they didn't work well. Now we have answered a lot of these
types of questions. Again, referring to the paper, these
are a set of techniques that we felt needed improvement.
It was sort of written by a committee, so we have all the
good things that the speech understanding people would like
to hear. From my point of view, we feel it is most important
to do signal conditioning and speaker variability. To us
that is the number one problem. If you are interested they
are all written up in this paper. One other thing I want to
stress again is performance evaluation. We have been burnt
in the past by insufficient performance evaluation. What I
would like to see is when you actually compare the three
speech understanding systems in terms of complexity.

Thank you.

1.11.2   Viewgraphs used by Dr. Beek are at Attachment 9.

1.12   Major Carlstrom noted that there was a tie-in between
the previous speaker and the efforts of the NATO working group
on speech understanding. He, therefore, introduced Dr. David
Hodg. from the U. S. Army Human Engineering Lab at Aberdeen
Proving Grounds, Maryland, as the next speaker.

1.12.1   Dr. Hodge's comments were as follows:

Bruno Beek has just given about half of my paper so this will be very short. Research Study Group 4 is organized under Panel 3 on Physics and Electronics organized under the civil side of NATO. The group has been operating since 1971. The topic is Automatic Pattern Recognition. There are seven participating countries. The primary objective of the NATO study groups is to look at, in this case, the military applications of automatic pattern recognition, to perform state of the art assessments on selected topics, and to find probable areas for cooperative research. The NATO mechanism provides a means for cooperating internationally on topics that are not appropriate in other media. I do not mean to imply that we only work on classified problems but it does provide us with a very simple basis for cooperating at least within the NATO community on problems that are difficult to cooperate on otherwise. The activities that we have been involved in include developing a system for organizing all the research that is going on in the various countries, in exchanging summary reports on the national projects (there are two summaries that are available from the Defense Documentation Center), identifying common areas of interest and, also, those areas in which there is no interest. As an example, there is no interest in cooperation on passive sonar because we would not only have to tell people what we know about Russian equipment, but we would also have to tell people what we know about English equipment. We don't want to do that. We are also conducting technology assessments on selected topics. We started with image processing. We have done one on speech recognition, and we have one on mechanical wave processing techniques which includes processing techniques but not data that are applicable to the acoustic side of sensing areas which is scheduled to begin in

February. We have developed one cooperative research
proposal in the area of image processing and the project
has been initiated. This slide outlines the steps we went
through in assessing the topic of speech recognition. We
developed a list of probable military applications that
we would like to automate. Bruno has already shown that
slide so I won't show it again. We had specialists from
the participating countries come in May of 1974 and present
their independent assessments of the state of the art,
present problems and the estimated cost of solution in the
probable system requirements for realization in the various
application areas. We have prepared the U. S. technology
assessment for that meeting. In August of 1974 we got
some expressions of interest in cooperation. There is
no defense supported research going on in Canada, Denmark
or the United Kingdom. The primary interest in cooperation
was expressed between the Netherlands and the United
States. Bruno Beek is preparing the final version of the
technology summary paper which he gave me this morning.
Therefore, it is not complete but we will shortly submit
it for publication as an unclassified NATO report. We
will discuss in February at our meeting the prospects for
cooperation on classified problems. We found no basis so
far for cooperation on unclassified problems as most inves-
tigators, particularly in Europe, indicate that almost all
the speech research is being done not by in-house organiza-
tions but on contract in the academic community. They feel
that existing information exchange media in international
conferences, journal publications, personal communications,
etc., provide an adequate basis for cooperation. In the
minutes of this meeting I will have the technology summary
tables that Bruno has already presented so I won't present
them again. We have a list of military tasks we would like
to automate. We have a list of techniques that have to be

perfected and a narrative description of each. We have
a statement of the state of the art for each one of these
techniques using three ratings: a - is useful now; b -
shows promise; and c - is a long way to go. Each one of
these things has been rated by the committee in all of the
participating countries. This table will be in the
minutes. Also, there is a list of the unsolved problems
and the list of unsolved problems that appears in the
minutes will be keyed to the processing techniques. Finally,
there are a number of near-term applications (you have al-
ready seen this slide) and these are applications which we
expect will be realized certainly within the next decade,
and in some cases conservatively estimated in the next
decade.

Thank you.

1.12.1   The vu-graphs used by Dr. Hodge and the tables
mentioned in his remarks are at Attachment 10.

1.13   Major Carlstrom recommended that everybody
take the opportunity to track the RSG-4 Papers. He noted
that he had had several of his people in his contractor
community request them and in some cases was able to get
them while in others he was not able to do so. Not only in
the speech area, he commented, but also in the imagery area
these papers are very useful.

           Major Carlstrom next called upon Mr. Jack Boehm
of the National Secruity Agency at Fort Meade to address the
group.

1.13.1   Mr. Boehm's remarks are as follows:

I have been asked to make a few brief remarks about
NSA's interest in speech research. Our agency has had an
interest in speech research for more than 10 years now, and
in those 10 years many of the problems we saw then are still
with us. The world that we live in looks like this. It is
a world all too familiar to many of you. The limitations
on bandwidth and signal/noise ratio are truly handicapped.
It is natural that those working in the research area to
automate voice processing are anxious to maximize their
chances for success. So, they intend to avoid these handi-
caps. Well, unfortunately, we have to live with them, and
systems designed to meet more ideal conditions often have
to go or undergo extensive revision in order to be useful
in this particular kind of environment. Now our interests
are broad, many of them overlap with those that Bruno Beek
has outlined for RADC previously. To sum them up, we are
interested in automating voice processing. Can you factor
speech into its components of words and get talker identity
into the language? We are also interested in seeking
efficient means of speech coding. We must minimize the cost
of storage and transmission for voice. Like RADC, we have
an interest in techniques which might enhance the intelli-
gibility of speech which is recorded under noisy conditions
or is distorted by one kind of a communication channel or
another. Our approach to these problems is always colored
by the particular environment that we live in. It is
necessarily constrained by conditions such as the band-
width and the noise limitations. I'll focus attention on
our word recognition interests because I can use this to
illustrate that point, and many of you, if not most of you,
are very familiar with the recognition work done on the
speech understanding project. Now this slide is a kind of
A/B comparison of speech understanding, and so is the
nearest approximation to speech understanding which might

be of practical use to our agency. On the left of the
slide is the sort of defining structure of speech under-
standing which is a fair representation lifted from the
Gould report. At the right, using the same structure defi-
nition, is the kind of a system which we might use. The
first comment we always get is that the constraints on the
right are almost uniformly more difficult. I suppose that's
true. What must be most appalling to those with artificial
intelligence orientation is the almost complete lack of
available syntactic and semantic support. These are areas
where many interesting questions arise. However, for us to
make use of such a system someone would have to produce what
Norm Chomsky might agree is a completely adequate grammar
for English, and other languages that may be of interest to
them. I think there are none of us in this room who can
see that happening in the reasonably near future. But still
I am optimistic in the sense that I think there are very
reasonable extensions of the state of the art that now
exists that can produce useful systems. The speech under-
standing project did make beginnings at recognizing words
as they occur in continuous speech. Also, I think we should
make a considerable effort to attack the multi-speaker
problem to see if there are practical ways to normalize for
talker differences. The notion of working with telephone
quality type speech, the bandwidth limitations that go with
it, the noise conditions, these must be tractable problems.
After all people do communicate under these conditions. I
think we should seek maximum advantage from phonological
constraints. Here again, some nice beginnings were made in
the speech understanding project. So, for the kind of word
recognition work that we might be interested in, the emphasis
would lie in those areas. I tried to illustrate those here.
There are also some other questions that might be examined
quite apart from trying to develop such a system directly.

Can we do something about the influence that regional accents have on word recognition. There seem to be sort of regular phonetic shifts. I don't know of much work that anyone has done to see to what extent you can compensate for those. You can study them quite apart from developing a system which will itself determine the probability of errors in these things. If you rely solely on phonetic types of information, one of the likelihoods of confusion is the phonetic strings occurring in the language. Another point would be, can we determine just what is an upper bound on a system which attempts to recognize words in continuous speech in the absence of syntatic and semantic support. That is really sort of it as a thumbnail sketch. We are forced to work with limitations, but our interests are fairly broad within those limitations. I, for one, am optimistic that there are reasonable extensions of the state of the art that can prove to be quite useful for us.

Thank you.

1.13.2 Transparencies used by Mr. Boehm are at Attachment 11.

1.14 Next to be introduced was Dr. Mundie of the Aerospace Medical Research Laboratory at Wright-Patterson Air Force Base, Dayton, Ohio.

1.14.1 Dr. Mundie presented the following comments:

We are actually a little bigger than Ear research. The program of interest to us, and from which we now speak to you, was a program organized about 1960 called the bionics program. Bionics is a name that sort of fell into disrepute so we don't use it anymore. But it was a phrase, a word coined by the people there at Patterson, to identify this area

51

of work. The area of work was defined as using a living system as a paradigm as an example for improving hardware. We still continue this and actually one of these word recognition systems, the one marketed by Tom Martin, was supported in the early development when he was still with RCA at Wright field under this bionics program. So we have a long-term interest, but we are aimed toward signal processing. In this bionics program let me say we are interested in two areas. We are interested in application of speech recognition technology. That particular area is being handled by the human engineering group at Wright-Patt rson in the Aero Medical Laboratory. They are interested in actually putting speech recognition systems to work in the Air Force and particularly in inflight application. We see the basic problem as being a natural language communication with computers, and for immediate application within the Air Force we sort of summarized it according to that particular chart. We see practical applications in the near term in heads up, hands off, utilization in the cockpit much as Commander Wherry talked about. We see, also, an inventory management within the Air Force. Communication with computer - the Air Force logistic system is all computerized at the present time - is done through hand-operated terminals. So we are interested in evaluating speech recognition systems and putting them into application, that is, the human engineering part of it. You have heard mentioned by the group, the ARPA group, by Bruno and Dave Hodge and others that there are problems in making the system work and these problems deal with signal processing, specifically feature extraction and that order of problem. This is where my particular interest lies, in the study of the auditory system. We claim that we are working with the paradigm of speech recognition. This is the benchmark against which all speech recognition systems are compared, namely

the human auditory system. We feel from our work and study
in how the ear functions that a great deal more information
can be extracted from the signal than is being extracted,
particularly with the systems that are in use in speech
understanding systems today. They all start off with band-
width limitations with the frequency domain analyses and
we feel that is an improper approach. From the very bottom
if you feel there is any virtue in copying a system that is
working you are starting wrong from the very beginning. So,
I would like to make that a point today. The auditory system
works in the time domain rather than in the frequency domain.
I will offer you a little evidence to substantiate that.
So, right off the bat, you are processing all your data in
the wrong domain. We are into the speech recognition
business because we have to test the hardware that we evolve
from our studies of the auditory system. The hardware as
it appears today, is outlined here; we are taking the sig-
nal through two to three transformations, the second and
third are a little vague, a little hard to separate. The
first significant transformation is a model of the mechani-
cal function of the inner ear, the cochlea model. We call
it the cochlea transmission line. This is producing 48 out-
put channels of analogue data and this data is being sampled
by models of the primary auditory neuron, which are basic-
ally signal-dependent encoding devices which have a pulse
train as an output and these are interlaced into a network
of dynamic controls so that actually the features that are
selected from the 48 analogue signals are signal-dependent.
The features of the system and the network adjusts itself
as time goes on. As the signal changes, I should say. The
net result of all this processing is a change, a transforma-
tion of the signal from a 2 dimensional amplitude versus
time into a multi-dimensional amplitude versus time trans-
formation, and then, from there, into multiple pulse trains.

The processing in the nervous system is carried on in these multiple pulse trains. So that incorporates the pulse processing, this section down through here, and this is an existing piece of hardware. The output of this system is 64 channels of pulse data which are multiplexed into 32 of computer interface here called the ASPP which only accepts 32 channels, so we switch between voices and voiced signals when we are processing the speech, 32 channels being devoted to the voice list and 32 to the automatic switching under the network control. This is flowing into the PDP-11 computer which is a final processing output device so we have 32 channels flowing into the computer. These are pulse channels. The computer is accepting these in parallel with the resolution of 5 micro-seconds on each of the channels. We measure pulse intervals on those channels to an accuracy of 5 micro-seconds. I thought perhaps our time would best be spent in talking about signal processing and how the ear functions in the time domain, so that is what the rest of this talk will be devoted to.

The first transformation that takes place is done in the inner ear by the cochlea. This is the transform from two dimensions to three, the dimension that I show here as distance is most often thought of as frequency and the ear is most often modeled as a set of band-pass filters. It is, in fact, a transmission line. It is a very unique sort of a transmission line. I will try to illustrate that to you. Speech, when input to the system, is transformed into this sort of a 3-D transformation. The first point I would like to make is that if you put in 2 sine waves - this is the sum of sine waves here - you see relatively poor separation. Those 2 sine waves are an octave apart and yet they are very poorly resolved along that dimension of distance which is frequently called the

54

frequency domain. That is the length of the transmission
line and in the ear it has a physical weight to it. It is
very finely resolved by the way. There are some 20 thousand
plus cells that are arranged along that length. So when you
look at speech signals, they are also distributed in that
dimension of distance. This is a picture of the vowel 'e'
and you can see very clearly high frequency or short interval
information in the foreground. There is some dispersion of
speech signal. The 'e', of course, being the extreme and that
is the widest separation of the two areas of interest along
that dimension. Let me talk about transmission lines for
just a moment for this is, indeed, the proper way to model
this transformation. It is a very unique sort of trans-
mission line. In a normal transmission line a signal is put
in on one end, it propagates down the line and goes out the
other end. Usually they are designed to not alter the
signal but to delay the time. This would be what we see if
we sort of froze the action for an instant. Along the normal
transmission line we see one or more cycles of a sine wave
stored in there. If we looked at it in the 3-dimensional
domain we would see this sort of a presentation where the
horizontal line depends on time rather than distance. Distance
is the verticle axis. You see that in any given instant in
in time a slice through here, which is what this is. There
would be multiple cycles or less than a cycle, depending on
the frequency stored in the line. In the inner ear this is
a very non-uniform line and the velocity of propagation
varies as to the function of the distance along the line.
So you see the wave peaks marked here with the dark dots,
how they curve indicates the case of the cochlea type line
because the further away it propagates down the line, the
slower it gets. There is another very unique relationship
in the ear and that is the fact that it is a leaky trans-
mission line. The signal leaks off of it as it propagates

down, it leaks off in such a way that there is always
exactly the same number of cycles stored in that trans-
mission line regardless of frequency. This particular line
is storing one and three-quarter cycles in it and before it
dissipates you see the signal start down the line and it
grows in amplitude and it dissipates very abruptly and be-
fore the dissipation there is up to two cycles stored in the
line. Here is a different frequency in the line 500 hertz,
again you see the number of cycles stored in the cochlea
line in contrast to what would be happening in a uniform
line. You can design these non-uniform lines to store
different amounts of signal and the amount of storage is
a feature which is generally ignored in cochlea vowels.
The 2-cycle storage is a piece of information that we came
upon by neurophysiological data measuring the response of
the nervous system and we measured these propagation
velocities from the responses of the nerve system and found
that the guinea pig here stores about two cycles. You can
design the line so, as far as engineering is concerned, it
can store different numbers. We think the fact that it
stores more than one and less than two is a very significant
fact in terms of signal processing in the ear. Back to
this illustration for a moment, we have looked at speech
for testing this particular system and we developed a
display device that can give us a real-time output of this
particular transformation of speech. We can write it out
rapidly and study this and we spent a number of years
studying the features of speech, after speech signal had been
transformed in this way, and we found for voice sound that
the most significant features, in terms of identifying the
voice segments, are what we call the first two intervals
and a pitch period. You see the periodic iteration of
this wave form. Each of these wave forms is produced by
one excercitation of the vocal track. We found that in

measuring, we are interested in two or three of what we
call first intervals in a pitch period. If you look in the
foreground you will see a short spacing between the peaks
and a background of wider spacing between the peaks.
That is what I mean by first interval - it is the interval
between the first two peaks that appear in each of those
pitch periods. So, in the 'e' we get two measures. We get
one in the foreground and one in the background. We have
what we call IA and IB (Interval A and Interval B). This
is sufficient information to identify that particular
position in the vowel space. So by extracting just those
two pieces of information you can identify it in the vowel
space. So what I am talking about here now is the ear's
capability to identify and classify each pitch period that
is produced by the vocal track and place it in position in
the vowel space. We have done some cycle acoustic experi-
ments that demonstrate that people can do this and, in fact,
you can classify and make the same errors you make with
sustained vowels on individual pitch periods. If we just
give you one of those things like a (sound demonstration)
you can identify it given the right experimental situation.
In fact, all that we have to give you is the first half of
a pitch period; you can classify it and identify it. This
graph shows you the vowels - these are the long vowels -
classified according to the sonograph analysis of placing
them in the form of one form with 2 plots and these are
the same signals classified in an Interval A or Interval I
and Interval 2 plot. Then you see a slight improvement in
separation of the groups in that space. This is for the short
vowels more tightly clustered and that is the frequency
analysis plot and this is the interval plot. What I want
to make clear is that we are at least as good as frequency
analysis and we think that there is a great deal more
information in there that can be extracted. Those

identifications are made from single pitch periods, not from
a sustained sample. A little bit of neurophysiology just to
demonstrate to you again that the auditory system is function-
ing in the time domain rather than in the frequency domain.
In the upper lefthand corner you see neural responses called
post stimulus time histogram or pulse occurrence histogram
which measure the activity of the single nerve cell that is
sampling this signal along the vaso membrane. This is the
analogue signal which is the motion predicted by the model
for the structure at that point. The bottom two illustrations
show where we have matched these two up and, as you see, what
happens is that there is a very precise encoding in the time
domain of the information. The nerve cell is following the
motion of the vaso membrane. I will just quickly illustrate
what some other signals look like. This is a sinusoid, two
different neurons, a sinusoidal input to the neuron and then
a speech input to the neuron. As we look at the single
neuron with different speech sounds from different vowels you
will see again that the neuron follows very faithfully the
motion of the vaso membrane as predicted by the model at
that point. Now, two things come from this: One is that
this, I think, is reasonable confirmation of the predicta-
bility of the model and the accuracy of our model in this
transmission line; and the second is that the neural infor-
mation at least after these two transformations, one the
transformation produced by the cochlea on the transmission
line, and the second, transformation of the encoding into
the pulse domain. It is still a time domain operation. So
the system, at least through its first two transformations,
is still functioning in the time domain and by making measures
in this domain we can identify the speech pattern. The
difference here, you see, you can easily transform from the
time domain to the frequency domain. The difference here
is in the order of magnitude, when you are transforming into

the frequency domain you are working with some time cycle
that is involved. You have to do your frequency analysis
over some time effort. The auditory system is not con-
strained to this. It is functioning in time and therefore
it can make instantaneous, and does make instantaneous,
measures. You see as the speech signal varies instantaneously
throughout, it could later on follow these time domain
changes. By studying lots of neurons you can develop some
sort of a composite picture like that and that's an illus-
tration of two different vowels looking at a dozen or so
neurons arranged along the dimension of the transmission line.
They all are following the particular wave form that would
be predicted by our model at that point. We want to test
these models and speech has long seemed to be a good test for
auditory system models. It is something that the auditory
system handles very well, does in real time, and makes
relatively few errors. We have been using speech as the
basic test vehicle for testing our models. This is to
illustrate to you the output from the computer that sits on
the end of the system, and it is receiving now as its input
multiple pulse trains, parallel pulse trains, there is
a plot here so that you see what are called F zero or pitch
changes being plotted here as interval lengths. It is also
doing some amplitude measures, the network is, and you are
seeing amplitude plotted here as a function of time in three
DB steps, this covering about 27 DB range here in this plot.
This dimension here is one line per pitch period, the com-
puter is doing an analysis and measures each pitch period.
This set of numbers over here is one of our present classi-
fication and identification schemes that we're looking at
for identifying the individual pitch periods. The numbers
represent measures that have been made on these pulse trains.
This fine-grained analysis of speech leads us to some features
that we haven't found described other places. You also can see

59

some fine-grain changes that you don't see in some other methods. For example, a little amplitude dip here during the 'v', you can detect it very well. A feature that I haven't found described before - perhaps June Shoup may correct me here - but the lengthening of the pitch period, or the lengthening of the interval prior to a transition from a voice to a voiceless phoneme for about a dozen pitch periods or 10 pitch periods or so, you get progressively longer and this is a little feature that cues you into the fact that all the sound is going to be a voiceless sound. So these measures are being fed to our computer system in real-time. This information is available to the computer in real-time. The extraction of the features and the classification by the computers are done in non-real-time. Basically we're attacking the problem then of improving the accuracy at the acoustic phonetic level. And the unit with which we work is the individual pitch periods so that phonemes are built up of sequences of pitch periods. Just to illustrate in more familiar terms, part of the information you can get out of this is a plot of interval versus pitch period number and this is what would amount to, I think, an F1, F2 tracking task which we've just plotted as a function of time. The length of the intervals that were measured is the first interval measurement that I was talking about. That is for void or avoid, here's a plot form, or wait. This kind of information, of course, is available to the computer and is supplied to the computer in real time.

In summary, these are some of the features of our automatic speech recognition system that we have here. The fricative identification is, essentially, a spectral analysis; it's done in the time domain but I think the features it was measuring are no different than you'd get from spectral information. All the voice sounds are identified with time

60

domain analysis and we're using only a few measures in each
pitch period to locate that position in the vowel space. The
digital computer task is much simplified by the amount of
information that we're supplying to it in these feature
extraction networks. Each pitch period is classified and
placed in the vowel space and we're working on how to separate
this into phonemes, obviously this is a dynamic thing---
the patterns move about and your tracking pattern motions
are generated by the speaker in the vowel space. We demon-
strated that there's very close relationship to performance
for the diffuse vowels but as you get to the more compact
vowels there's not this one-to-one correspondence between
interval and frequency that we can see in the more diffused
vowels. The pattern space, as far as the computer is concerned,
is a set of intervals or measures that would be extracted
from pulse trains. As a matter of comparison, we ran a test
with normal vowels and nasalized vowels using five pitch
period exerpts from continuous speech record to a con-
tinued speech utterance and took out five pitch periods and
used that as a test signal. We gave those to a panel of
listeners and to the machine. The panel of listeners had to
then identify the vowel. That was their task and we found
that there was about 50% accuracy on placing the vowel exactly.
If the vowels are nasalized that drops to, I believe, 33% or 35%,
the machine had comparable measures of 46% and 31%. If you
included the nearest neighbor in the vowel space then the
percentage is increased tremendously. This was precisely
on target and this is talking about the nearest neighbor in
the vowel space. The panel performance jumped to 79% and 72%
and the machine jumps to 73% and 64%. So, I think we'd like
to claim that you can get a lot more information out of the
signal prior to operating on your phonetic identification
and your word identification and your syllable separation.

You are really penalizing yourself, throwing away a tremendous amount of information that's in that signal. You're ignoring it and not using it.

Thank you.

1.14.2   Illustrations used by Dr. Mundie we not available for inclusion in these minutes.

1.15   Major Carlstrom pointed out the fact that Dr. Mundie actually operates a wet laboratory to collect empirical data and formulates it into various network arrays for processing.

He next introduced Dr. Donald Christy from the Naval Electronics Laboratory, San Diego, California.

1.15.1   Dr. Christy's presentation follows:

The principal issue I want to address are the aspects of speech processing that we will have to address in the future if we want to make them applicable to military environments in the field. I am going to talk about two things, both of them do not apply to speech processing, per se, but they have to do with this process of trying to return the cosmic things into the work so that they can be used. The first one is in regard to the use of micro-processors, essentially to process natural language with the idea that this would be extended to the areas of speech processing when it could be shown it could be done with natural language. The first problem is to, essentially, parse English with a micro-processor supplemented with a disc. We are using in the system a diablo disc 44 connected through a cache memory used to provide a buffering both in time and access. We're using an intel 8080 processor connected with a teletype-writer for the purpose of inputting the English text. I'm

talking here primarily about the parsing process. There's
also some work going on that has to do with the processing
after we get a parsed structure and process it to accomplish
the semantics. In principal this is an information retrieval
system. The approach that I'm using in the implementation
on the 8080 is to use Martin Code's type positive technique.
This is a non-deterministic type of parsing and since they
are built up a tag rather than a tree, for obvious reasons,
then it also has capability to handle additional pieces of
information in the austere type of grammar. We use features
in addition to the regular type of parsing structures. The
features are essentially tags which can be tacked on to each
positive step and then transferred forward along with the
parsing structure. In addition to that, it's easy to extend
this to include probabilities in other types of language that
you wish to consider. For the purpose of using the 8080
we would like to have the parsing done in near real-time in
order to accomplish this. It is not possible to wait until
the end of the input stream in the sentence, so the parsing
has to take place almost immediately at the start upon the re-
ception of the first character. Since it's an 8080 we don't
expect to do anything else except the parsing while we're
going along. It's not a time-share type of system. So the
problem is twofold. One is to devise and recognize the
particular patterns, if you will, and also to retrieve that
portion of the syntactic and lexicon that will be needed at
each step. We use a suffix identifier tree in order to
facilitate the syntactic rules and the lexicon. They are
bent into each other and we use the suffix rule because in
this manner we can essentially start the parsing process by
looking back at what has already arrived. However, in order
to supplement this, in order to retrieve information or
rules, we have to also look at the associate prefix at each

63

step and say, "What rules may start at this point?", and
then pull those in from memory rather than to have the
entire lexicon and parsing rules in the memory at the time
the system begins. In addition we are required to do some
amount of trimming because as rules are coming in some of
them will not be used, or some will have been used but are
no longer needed. In this case, when a new rule comes in
you have certain sets of structures that say that at the
moment the last set of parsing gets done you will have
knowledge about all the rules that are possible as you go
through the rules that you have already in the memory, then
you can prune out those that do not have parts just prior to
pulling in the next set of rules. This program is not very
far along in the sense that it only started a couple of months
ago so I can't tell you much more than that. The second prob-
lem has to do with the implementation of some of the parameter
processing in a low-power environment and the question is how
to do this with low-power. One of the techniques that might
be suggested is the use of optical processing. One technique
that we are looking at at the NELC is the work of Keith Brom-
ley, who is not related directly with my work but I felt that
it was significant so I wish to present it. The process is
to essentially do some parametrication and unfortunately we
haven't really been able to see how we could do LPC coding
but we have been able to do a little bit about other types of
things like variant analysis. The process associate has a
light emitting diode and it is modulated with the incoming
signal. To modulate the signal, and then to have a mask, the
mask and shadows are the modulated signal and then this is
coordinated with a charge-couple device type of thing that is
storing and deliberating the output. First the light comes in
at this point, it moves to the shadow through parts of this
area here. At each time interval the charge is moved through
the sliders to one step beyond and in that way, at the end,

64

the results of the integration over various steps allows
for the transform to be displayed in a register at this
point. Now to give you a couple of ideas as to what this
might look like: The mask is a critical item here, the mask
and the timing. This is the mask of a Fourier transformer,
a cosine transform. We have various other transforms
that we have masks for. You'll notice that because it's a
an optical system, you have to display negative and positive
parts at separate pieces of graphs, so you should notice
that the righthand and the lefthand slide in. This mask re-
presents the positive and the negative parts of the project.
Now this is another one which may or may not have usefulness
in speech. This is a Walsh transform - it's a little bit
more uniform than the Fourier transform. There are quite
a number of transforms possible with this type of mechanism.
In fact, you can do matrix multiplication. This is an
indication of that. With matrix multiplication you can do
the Fourier transform but you can also do such things as
clustering. So the technique can be used in several ways
in speech processing and this is what our intent is - to
look at it in those terms. Present research is going on in
the process of trying to allow us to change the matrix part
of the mask dynamically. This will allow possibly for such
things as LPC coding but until that is done we can't really
accomplish LPC coding. We still only have the input and the
output available. Inverse matrix can also take place with-
out too much trouble. O. K.

Just to conclude, I am going to show a couple of
pictures of some of the equipment. This is a laboratory
set-up. There is the modulator on the left and it goes
through and passes through the mask, goes over to the charge-
couple device at this end. We've been investigating several
techniques to remove the optical cylinder that is necessary
for doing this kind of process and it consists of using

65

instead of a diode an electrical luminescent type of
panel and then with that we can accomplish a sandwich nearly
a half inch thick rather than the optical path indicated
here. And with that I conclude my remarks.

Thank you.

1.15.2   Transparencies used by Dr. Christy in his talk
and some additional materials furnished by the NELC are
at Attachment 12.

1.16   Major Carlstrom announced that Dr. Christy's
talk would conclude the morning session which was then
adjourned for lunch. The group would reconvene for the
afternoon session at 1:00 p. m.

2.0   AFTERNOON SESSION

Major Carlstrom reconvened the workshop group at
1:15 p. m. and introduced as the first speaker for the
afternoon session Donald C. Lokerson from Goddard Space
Flight Center, Greenbelt, Maryland.

2.1   Mr. Lokerson gave the following presentation:

Introduction. Since Alexander Graham Bell began
studying speech in detail about 1866, thousands of re-
searchers have been frustrated in their attempts to unlock
the key to reliable connected-speech decoding. Just as
Bell's research in speech began by working with the deaf,
so this present work began, with the idea of building a
hand-held "calculator" which would display speech phoneti-
cally for the deaf. The approach taken in the process
parallels concepts used in some spacecraft telemetry signal
coding and decoding systems.

What is the Human "Channel Coder"? We form
vowel sounds by moving our lips, tongue and mouth into
various shapes, as shown in Figure 1. We form consonants
by making high frequency noises and nasal effects. All
vowels and some consonants include vocal cord vibrations,
making harmonics which resonate in the various mouth
cavities. People's lips can move at least 5 hertz per second,
but the tip of the tongue moves somewhat slower and the back
of the tongue moves only about one hertz per second. These
slow movements slur speech from one speech segment to the
next. This characteristic has made decoding speech seem
difficult, if not impossible. The frequency spectrums of
speech are very complex, variable, and highly dependent
upon the person talking. It becomes clear that this is not
the real key to decoding speech. Just detecting the major
speech components is equally unsatisfying. Any "channel

67

coder" needs to be composed of symbols. In speech, these
have been called phonemes. The spacing and separation of
phoneme symbols needs to be wide enough to prevent confu-
sion for reliable decoding by the brain. It seems probable
that we learn to speak by moving our mouth organs to make
these symbols, using our ears to act as a feedback system
to "zero in" on the symbols needed. This is confirmed by
the problems post-lingually deaf people develop as time
progresses. We know that speech can be distorted in almost
every conceivable way and still be understandable because
of the redundancy built into the coding process. For instance,
speech over a telephone is limited to a 3 hilohertz range,
and is understandable even under such conditions. Since
many uses of speech decoders should work over such conditions
we will limit our considerations to this "channel". For
noise immunity, "touch-tone" telephone systems use pulses
of two frequencies from about 700 hertz to about 2700 hertz
to "speak" number symbols. It appears that all speech can
be decoded much the way "touch-tone" symbols are decoded,
with some important differences.

A child's mouth organs are a different size from a
woman's and a man's mouth organs are bigger than a woman's,
proportional to their head sizes, as shown in Figure 2. This
makes the vowel frequencies generated different, in a system-
atic way, as shown in Figure 3. We will show that this size
difference can be compensated for by taking the ratio of the
second "formant" with respect to the first formant, and the
second formant with respect to the third formant. These
formants are resonant points created by the mouth organs:
some examples are shown in Figure 4. Speed variations in
disk or tape recording result in intelligible speech over
quite a range also. The vowel phonemes of speech can be
plotted into such a code-symbol set into a "phoneme space",

as shown in Figure 5. The separation of these symbols is improved and independent of the speaker. Data by Peterson and Barney[1] give a confusion-matrix of English vowels. That is, when one vowel is spoken, a percentage of occurrences are mistakenly perceived as a nearby vowel. In Figure 5, notice the reasonable correlation between the human results of the phoneme space.

The Vocal Cord Modulator. The vocal cords vibrate with air flowing through these muscles. As shown in Figure 6, at the top, the waveform is triangularly shaped, and thus is rich in all harmonics. A man's voice changes pitch by at least an octave (factor of two), changing in repetition rate more than in wave shape. Women and children have higher frequencies of glottal waveforms which can reach up to above one kilohertz when singing. The middle portion of Figure 6 shows various harmonics needed to make two different vowels. In the example shown, both vowels are composed of the same two harmonics for the female vocal cord harmonic. The "E" has more of the 300 hertz component while the "ae" has more of the 600 hertz component. Notice that zero crossing detectors would get the same result and would not distinguish these differences. The equivalent male waveforms are different. This corresponds to the practical results in which men's voices can be decoded better than women's or children's. This is because the man's harmonics are closer to each other, and thus define the resonances better. In practice, however, a woman's voice is as easily understood as a man's. This means that our hearing processes probably have a different way of detecting at least the first formant as discussed below.

The Quantizer - A Key to Speech Decoding. As the waveforms at the bottom of Figure 6 show, some method of detecting not just hertz-per-second is needed but a method which is proportional to the amplitude as well as the

69

frequency of the complex waveforms. For the female case, we would want a higher value for "E" than "ae". The top of Figure 7 shows one way of achieving the desired result. When the input waveform is above its average no output pulse will be generated. As the waveform goes below "zero", a pulse will be created at the output. If the pulse goes even more negative, the original pulse is inhibited. This can be implemented by a simple circuit, shown at the bottom of Figure 6. If the input waveform goes even more negative, another pulse will be generated. If the pulse goes even more negative, this pulse can be inhibited. As the waveform progresses back toward "zero" the reverse operation occurs. Thus in this case four pulses can be produced for each cycle. Figure 8 shows five different cases of possible speech waveforms and their corresponding quantizer output pulses. Note that these are shown with the very special phase relationships created by the glottal waveform. Figure 9 shows how the first formant mouth resonances might look for various vowels in the frequency domain, and under the same conditions as the previous Figure. Note that this technique produces pulses proportional to the center of resonance even when the glottal waveform harmonics straddle the center of resonance. Using this concept, it is easy to see that women's and children's voices may be decodable as easily as men's. The technique can be expanded to include both positive and negative portions of waveforms. It can be made into an algorithm such as a pulse being generated for even-numbered millivolts but not for odd-numbered millivolts, possibly generating hundreds of pulses each input cycle. This process is probably analogous to the ear-brain operation in this way. The cochlea nerves fire proportionally to the amplitude of the frequency to which each is sensitive, thus the brain does some correlation similar to making pulses proportional to frequency.

70

For the quantizer to operate properly, it needs to have a normalized output independent of the volume of the detected speech. An automatic-gain-control can accomplish this.

Vowel Discrimination. Figure 10 shows the bandwidths of vowel resonances as depicted by each vowel symbol. There are three of each symbol to represent the three formant frequencies. The bandwidths are only about ± 40 hertz for the first formant, and a maximum of ± 200 hertz for the third formant. This means that the harmonics of the glottal waveform will not be likely to be centered in the mouth resonance. The dots in the graph at the bottom of Figure 10 represent the differences between frequencies of English vowels. Notice that they are about equal to twice the bandwidth of the resonance. Thus the value of the quantizer to determine accurately the resonance point becomes clear and very important, and could be of value for all three formants, but particularly for the first formant.

Consonant Discrimination. In the English language, the consonants are made up of plosives, fricatives and nasals plus some vowel-like sounds. Most researchers attempt to detect the high frequencies of fricatives such as "S" and "SH" as shown at the top of Figure 11. However, some of these do not pass over a telephone link and yet can be understood. It is true that the speaker makes these noises by putting the tongue in particular shapes which affect the vowels which precede and follow the consonants, particularly bending the second and third formants as shown at the bottom of Figure 11. Three of these are voiced and three are unvoiced, and are probably distinguished from each other in that way. Figure 12 shows the equivalent results for plosives and nasals. These shifts in vowel characteristics may be used to detect the consonants by modified areas of the phoneme space, as shown in Figure 13, with about 30

71

phonemes shown. Others can be defined as more information
is gained on them.

How the Hardware Works. A block diagram of the
analog portion of the decoder is shown in Figure 14. A
microphone picks up the speech which is fed to an automatic-
gain-control amplifier. The amplifier normalizes the average
signal level and background noise. A pre-emphasis filter
compensates for the decreased spectral energy at higher
frequencies. The diodes represent a possible way of empha-
sizing the glottal ringing by giving a logarithmic gain with
respect to voltage. The fast automatic-gain-control amplifier
follows the rapid changes in speech amplitude. The capacitor
by-pass gives the unit a rapid roll-off above about 3 kilohertz.
The Zener diodes give a possible logarithmic gain with respect
to signal amplitude to emphasize the glottal waveform ringing.
The amplifier gains can be controlled by the AGC buss from
several points. The first formant would give good normaliza-
tion for the quantizer. The second formant could be used so
that the quantizer would produce pulses partly proportional
to the strength of the second formant amplitude. This would
be less affected by the glottal vocal signal. The third
formant may be needed to give good separation for fricatives.
The optimum arrangement has not been determined yet. The vocal
cord filter passes the glottal waveform to a threshold detec-
tor to determine the voiced or unvoiced characteristic. The
first formant filter passes the spectrum mainly between 300
hertz and 800 hertz. The quantizer is at the output of this
filter. These pulses are counted by the "F1 counter" until
64 counts are detected in the second formant or until 60 milli-
seconds have been counted. The second formant filter covers
about 900 hertz to about 2500 hertz. The output is detected
by a threshold and Schmitt trigger circuit so that only the
strongest components are detected. The output is a series
of pulses proportional to the second formant. The third

formant covers the range from about 2500 hertz to 3500 hertz,
and operates much as the second formant filter does. The
pulses are counted in the same way as the first formant
counter. A fourth formant has been shown as a possible aid
in detection of the telephone ringing, sirens sounding and
possibly for fricative detection, but sample rates would
have to be faster for such purposes.

The outputs of the counters go to Figure 15. There,
buffers hold the data for display purposes. An oscilloscope
displays the two axes which have been normalized for mouth
size. People who have used the device agree a strong biofeed-
back exists as one speaks to direct one's voice to areas of
the screen. It appears that this display would be very help-
ful to train deaf people to speak. The digital counts go
from the buffers to digital-to-analog converters to drive
the oscilloscope. The digital outputs can also drive
"programmable read-only-memories" which provide the "table-
look-up" feature to segment the speech into phonemes. After
the normalization process which compensates for mouth size
and the quantizer process, the resulting output makes a
"phoneme space" which appears to uniquely define the phoneme
spoken. The output counts of the two counters form an address.
At addresses which define phonemes, digital contents are
stored to define the output desired. This could be a light-
emitting diode display code, a computer input code, or any
other digital symbol. Much of the phoneme space is "blank"
and does not represent a decodable phoneme. An integrated
circuit containing the table-look-up is shown in Figure 16.
It takes 8 bits of input addresses and reads out up to 8 bits
at each address. A buffer storage may be used to insure
that valid phonemes were detected by two adjacent samples.
The buffer could also employ rules of spelling and s_ntax
to convert from phonetic spelling to more conventional English,
before displaying the information. The table-look-up may

73

take a form shown on Figure 17, for voiced indication. The
detection of "g", "d", and "b" may be possible to be separate
from the vowel associated with it, but experience will tell
better the degree of success this may hold. Detection of
vowels with the consonants may be more difficult but still
quite possible. Figure 18 shows the first demonstration of
the unit which employs the techniques described in this
paper. The 21-inch oscilloscope in the center of the photo-
graph displays the dot which moves to the three points shown.
We decoded the vowels in "team", "tin", and "says" reliably
for my voice, my wife's voice and for those of my daughters,
ages 6 and 9. An LED character at the top right showed the
characters "[", "I", and "E" respectively for the vowels.
The analog hardware is on one board to the right and the
digital counters are on the second board. The table-look-up
is in the box below the oscilloscope. The unit held in my
hand is a display unit which serially displayed the characters.
The unit was operated after only four days of checkout and
and setting up.

Potential Uses of the Speech Decoder. The original
concept of the unit was to be in the form of a hand-held cal-
culator with an alphabetic display and with the keyboard in
phonemes so the unit could synthesize speech for pre-lingually
deaf people. Another novel form of the unit could be a wear-
able device, shown in Figure 19. The eyeglasses are equipped
with an alphabetic display which makes a virtual image in front
of the wearer with the user speed-reading the conversation.
This unit would be inconspicuous in use and would be best for
post-lingually deaf people. It could also respond to the
user's voice and thus give him the feedback he needs to talk
well, even into old age. Detection of door knocking and tele-
phone ringing is equally possible to aid in the use of the
device by these same people.

74

The project was originally developed for deaf people since they could benefit even from an imperfect system. However, the new concepts discovered during this development appear to make the decoder reliable enough that many other uses appear quite possible. Unlike other devices already available, this unit needs no "training", and thus will work in conversations. Its noise immunity appears to be good. Thus it could probably do jobs such as court recording, dictation, and automatic equipment control by direct voice interfacing. For transcontinental communications, the unit could send the phonemes and voice pitch at much reduced bits-per-second, probably about 100 bits-per-second, compared to the present 80,000 bits-per-second. The bits could be scrambled for security and at the remote end a voice synthesizer working in essentially the reverse of this encoder would produce a natural sounding voice. It might be possible to achieve very natural voices which duplicate individuality well. This technique could save the expense of entire communications systems.

How Will the Unit Decode Other Languages? The vowels of all the languages of the world differ. A preliminary look indicates that languages such as Arabic and Eskimo have few vowels. Figure 20 attempts to show the vowels somewhat related to their method of generation, and hence somewhat related to the phoneme space which will result. Generally, the vowels are somewhat spread about evenly. For example, no one language uses only front vowels. Despite the varying degrees to which the different languages of the world have been developed, it appears that the vowels of these languages should be as decodable as those of the English language. Figure 21 shows data from a paper by Fant[2] for Swedish. Note that the table-look-up is different than for English, but not necessarily more difficult to detect.

Figure 22 attempts to show the consonants of the languages of the world. The table is plotted with respect to the part of the mouth which makes the speech, thus generally where the tongue is constricting. The bold characters are English while the light ones are used by other languages. Most of these appear to have the characteristics similar to English and thus should be decodable. It is unclear that the 38 clicking sounds of Zulu would be decodable, but this should not be ruled out. Some languages appear to use more nasals, such as Eskimo. Some inhale as well as exhale and this distinction may be hard to detect. Thus it is difficult to predict the problems which may be encountered in some foreign languages without more detailed study.

Conclusions. This work describes three concepts which are believed new:

1. The quantizer more exactly defines the mouth resonances.

2. The ratio of the first formant to the second formant and third formant ratio to the second formant appear to give improved decodability without the use of training.

3. The table-look-up techniqe allows an easy way to segment speech and convert speech to any arbitrary code.

The combination of these three concepts and space-age technology should make a speech decoder that is small, inexpensive, reliable, and thus available for a wide range of uses. These include aid to the deaf, vocal machine control, and communications. These concepts may alter a wide range of techniques used in speech analysis, pathology and related areas.

support. To Dr. Orin Cornett of Gallaudet College, I want
to express my thanks for his counsel. I want to thank my
wife, Judy, for her special ability to deal with phonics,
spelling and frequently analysis. My wife and daughters,
Mary and Susan, served as excellent examples of woman and
children for many hours for this work, and they put up with
my many hours at home and away from home to pursue this work.

<u>References</u>.

1. Gordon E. Peterson and Harold L. Barney,
   "Control Methods Used In a Study of the Vowels".
   Bell Telephone Laboratories, Inc., Murray Hill,
   New Jersey (1951).

2. G. Fant, <u>Speech Sounds and Features,</u> MIT Press,
   1973.

2.1.1   Illustrations and Figures referred to by Mr. Lokerson
in his presentation are at Attachment 13.

2.2.   Following the NASA Presentation, Major Carlstrom
called upon Mr. William P. Dattilo, the project manager for the
Army Tactical Data Systems (ARTADS) Project at the Army Mat-
erial Command, Fort Monmouth, New Jersey.

2.2.1   Mr. Dattilo described the word recognition for Army
Tactical Data Systems as follows:

Introduction - The Project Manager for Army Tactical
Data Systems (ARTADS) is tasked with the life cycle management
of tactical systems which rely heavily on source data automa-
tion devices. A number of hand-held message entry devices
have been developed and tested for accurate entry of tactical
data, yet none have demonstrated a completely satisfactory
combination of size, weight, cost, and human factors character-
istics. While continuing to pursue hand-held devices

as the primary method of messages entry, ARTADS has init-
iated a program of word recognition to determine the suit-
ability of word recognition systems for field use. This paper
will describe the ARTADS Word Recognition Systems (WRS)
and will address five basis areas definitions, specific ap-
plications, system configuration, program goals, and status.

Definitions - Word recognition systems operate under
three simplifying constraints: first, a gap or pause in speech
is required between each spoken word or phase; second, the
number of words in the vocabulary is limited; and third, the
system is individually trained for each word in the vocabulary
by the speaker; that is, the system is speaker dependent to
provide a degree of security. The system can also be made to
be speaker independent by adjusting the discrimination thres-
hold level to permit addressing by a multiplicity of users
without prior speech training of the system. Figure 1 shows
the typical operation of a word recognition system. During
the training mode, each word of the vocabulary is spoken and
undergoes an Analog to Digital (A) conversion and compression.
The resultant pattern is stored in the system memory. Once
the training has been completed the system will accept any
word in the vocabulary, will determine which word is spoken,
repeat the word back to the speaker for confirmation, if re-
quired, and/or speak the next field name in a message to prompt
the user utilizing a speech synthesizer programmed for what-
ever words are required by a specific application.

With the three stated constraints, the present state-
of-the-art for accuracies greater than 95% is a 30 to 100 word
vocabulary. For vocabularies less than 30 words, 99% accura-
cies have been demonstrated by a number of systems. The re-
quired word gap is 100 to 200 milliseconds.

Applications - WRS is being developed for test
using formats and vocabularies from the Tactical Fire Direction
System (TACFIRE), and the Tactical Operations System Operable
Segment (TOS^2). TACFIRE is a field artillery command and
control system which enables Forward Observers (FOS) to call
for fire on artillery targets. The FO carries a device which
enables him to enter sixteen different thirty character mes-
sages. A synopsis of these messages is given in Figure 2.
As shown, the preponderance of words spoken to enter a
TACFIRE message are digits. TOS^2 is a data storage and re-
trieval system and has a vocabulary of less than one hundred
words and five message types. The vocabulary for TOS^2 is
potentially larger than the TACFIRE vocabulary: however, as
an engineering tradeoff, certain fields in the messages having
over one hundred possible unique words were treated as a
coded three-digit number. The present contract calls for
the delivery of a system which implements both of these ap-
plications.

With vocabularies of the size of TACFIRE, training
is cumbersome. For the present effort, a straightforward
approach will be taken until the basic accuracy and utility
of the system is determined. If the accuracy proves suitable
the training problem will be addressed, taking into account
the considerations shown in Figure 3. One solution which is
immediately apparent is the treatment of the digits as
speaker dependent and the remainder of the vocabulary as
independent, or independent by speaker class. The latter
solution would incorporate a number, ten perhaps, of master
training sets against which each FO would be tested and
categorized or rejected as a usable speaker.

In addition to the applications under contract,
word recognition has been identified as a method of display

control, communications control, processor control, and
numerous other peripheral equipment which are under control
by a digital interface.

PMO, ARTADS, has under development a numer of com-
plex tactical situation displays which are ideal applica-
tions for voice control using word recognition. Other ap-
plications are being pursued with TRADOC, user schools,
test agencies, and project managers to establish requirements
for a LOA or ROC.

System Configuration - The WRS is under development
by Scope Electronics Incorporated, and the major specification
requirements are shown in Figure 4. The vocabulary of the
WRS totals 350 words; however, due to the structure, or syntex,
of the applications, the number of words required to be re-
cognized at any given utterance does not exceed 36 words.
The accuracy requirement for the WRS is 95%, demonstrated over
FM communications links with a 10 db signal to noise (S/N)
ratio. Preliminary WRS results with tactical FM radios and
handsets have been good, and prior to award tests were per-
formed using a system developed by the US Army Electronics
Command, demonstrating the feasibility of operation over FM
nets, clear and encrypted.

The block diagram of the WRS is shown in Figure 5.
The system consists of a three channel preprocessor with a
voice generation unit for each channel, a processor and
memory, a disk pack, magnetic tape unit, display and keyboard,
printer, paper tape reader, and card reader. Each channel
interfaces with a AN/PRC-77 or AN/VRC-46, the tactical FM
radios used by the FO with or without a security device. A
display operator with complete override capability is pro-
vided to monitor transmissions on the nets. As messages are

80

being received on a net, the translated data is simultaneously displayed in the appropriate area of the display.

The general operation is as follows: the FO enters a word into the system over the net, the WRS translates the word, displays the word, and transmits the translation back to the FO for verification or correction using the voice generation unit. Translation can occur on all three nets simultaneously. A single channel example is given Figure 6. In the example "ALPHA CHARLIE" (AC), "THIS IS", and "BRAVO ONE" (BI) are all treated as one word by the system. The USER column indicates the words spoken by the FO and the WRS column indicates the WRS reply as generated by the system. Transmission is initiated by the user and consists of three words: "AC" "THIS IS" "Bl". The WRS decodes "Bl" from an active user list, loads the vocabulary of the individual Bl from the disk, and then responds "Bl" "THIS IS" "AC". The second transmission from the user indicates one of sixteen possible message formats, in this case "FIRE MISSION GRID" (one word). The WRS receives the message type and from internal tables finds the first required parameter of this particular message, "NORTHING". The WRS replies "FIRE MISSION GRID". "NORTHING". The user then enters the northing data which the WRS has just requested. Transmissions occur in this interactive prompting fashion until the message is completed. Note that if the user falls out of syntax, speaks the control word "OPERATOR", or is identified as an imposter, the WRS automatically switches the display operator into the net. The FO has the option of correcting the data received by the WRS by using the control word "CORRECTION" as shown in the example.

The method of handling three nets simultaneously within the WRS is shown graphically in Figure 7. It

81

represents an extension of the single channel case making use of the disk capability in the system. The WRS stores training data for up to 64 speakers and allows sixteen to be identified as active users at any one time. These sixteen speakers can enter data on any one of the three available nets.

Program Goals - The goal of the program is to determine the suitability of word recognition for field use. In this contex, word recognition has an inherent advantage and an inherent disadvantage. The advantage is mobility. The WRS will enable remote personnel to enter date using no device other than the presently carried man-pack radio. The disadvantage is communications exposure. Although WRS should take no longer to enter a message than the present manual system, burst transmission is required to approach the transmission times exhibited by the hand-held devices.

The WRS will undergo acceptance testing at the contractor's plant to assure that the accuracy requirements stated in the specification are met. Subsequent to those tests, WRS will be moved to Ft Hood, Texas for tests with military personnel. The elements of the tests are given in Figure 8. Whether word recognition is successful for field use is primarily dependent upon its accuracy and cost with respect to the hand-held devices.

Status - The WRS hardware has been assembled and is capable of processing, displaying and printing the TACFIRE messages. A three-channel FM system is operational and the TOS software design is continuing. The system is scheduled for delivery in July 1976. Testing at Ft Hood is scheduled for August 1976.

82

2.2.2   Copies of transparencies utilized by Mr. Dattilo
are appended as attachment 14.

2.3   Major Carlstrom commented that there were two
scenarios. The first one is where you are trying to talk
during a scenaro with cannon fire. The second one though,
he indicated, is a problem. That is where the speech system
has to maintian channels for long periods of time. This is
worrisome because one does not want to compromise location
or capabilities. So with a letter key-type device you can
store up the information and then transmit by burst. In
this way one decreases exposure to site location or to tech-
niques.

Major Carlstrom introduced as the next speaker
Dr. Goldstein, from the Naval Training and Equipment Center,
Human Factors Laboratory, Orlando, Florida.

2.3.1   Doctor Goldstein's comments were as follows:

My name is Ira Goldstein and with me is Robert
Breaux and we're with the Naval Training Equipment Center
in Orlando, Florida - specifically with the Human Factors
Laboratory. Now the Naval T.E.C., as some of your may know,
and others may not know, is principally in the business of
producing large scale simulators for training. This is ac-
complished for such things as flight instruction, tactical
operations, rehersals and things of that sort. So we are
concerned with training. The Human Factors Laboratory in-
cludes a group of 21 psychologists who contribute to the
design of these simulators from the point of view of train-
ing concepts, instructional system design, and things of
that sort. Among the jobs that we see while looking around
the Navy is a class of activity that involves the use of a
highly restricted arbitrary language; a highly stylized

speech. Therefore, our concern is in developing training systems which permits us to take students and instruct them in the use of this vocabulary and syntax for particular kinds of jobs. Typical of these are; traffic control, ground controlled approach, and air intercept control. These are jobs where they have to limit themselves to a particular domain of discourse. One thing we're not in is in the speech understanding research business. From my point of view whatever it takes to understand what is being said is the given. We want to capture what the student has said, and build on that, because the kind of system and situations we're concerned with usually involve one or more large scale computer. With that kind of an environment you have the opportunity to build automated adaptive training systems. In order to do that you have to be able to objectively assess the performance of the students. So the function here is to capture the speech, compare it to some ideal model of behavior (what the person should be saying in a particular situation), develop some scoring system, and an adaptive form of sylabus that can assure progress of the student through the course of instruction. In that way, we would hope to exercise closer, more precise, control over the training process. Hoping to make a distinction here, many of the other applications that have been referred to involve substituting speech for things that are done in other ways today; for instance, with keyboards or some other form of perfecting or contributing to improvements in command and control. In our particular situation, speech is the very thing that we are concerned with,not substitution devices. We're interested in training people to talk as required in particular situations. For this purpose, we found that isolated word recognizers appear to be adequate for this kind of a category 1 applications. I'll pick up on what Dr. Beck said a little earlier about the fact that some of

these devices are quite adequate for many kinds of applications. And perhaps this group later in the afternoon, could consider the direction of effort and which way things could go in terms of producing an early success that would encourage support for speech understanding research. My personal opinion is that we can achieve a very visible kind of success by using a IWR in some organization. I will now introduce Dr. Krough who will tell you a little bit about the nature of the application: of the hardware sorting we have; and the training system that is being developed.

Dr. Breaux: In the spirit of the Newall Report, these are the dimensions of the speech understanding system which we use to train our traffic controllers in the precision of the approach phase of ground controlled radar approaches. Let me remind you that the speech understanding component is only one component in an overall training system. We're interested in training novice controllers (air controllers at the Navy's school for enlisted personnel) in Billington, Tennessee, in which both female and male enlisted personnel are trained. They're right out of boot camp, about ten weeks out of boot camp, and they know very little about the RT. They must be trained how to speak. They must be convinced that no pilot wants an excited controller. They must learn to say things in a systematic way. Everything the radar controller says, must be systematic regardless of whether or not the pilot is in a nosedive or is going to miss the approach. Given that, we want to use speech understanding as a simple component, therefore, it has to have these particular characteristics. We want to use it in our simulated aircraft approaches; that is, simulating an aircraft, and a pilot, and various kinds of wind conditions so that the controller is making non-continuous kinds of voices; "approaching glide path", "begin to spin",

85

"slightly above glide path", "coming down", etc. We have double buffering in our software in the sense that he may say two phrases, or three or four in rapid succession. "Coming up", "on glide path", "going above glide path", "slightly above glide path". Those four phrases, have been tried in our system and do work. While it's processing one phrase, it's including a second; that's how the double buffering works. The hardware we use is a "threshold" technology; VRP100, with a NOVA 1200 computer, and 32K. The "1200" is relatively small. A V0-track V6 is used; for speech synthesis, to communicate, to serve as a prompter to the student, to serve as an error feedback when he makes mistakes, and to serve to simulate a pilot's voice in trouble and an emergency situation and what have you. We need multiple speakers. These speakers are sequentially balanced. One speaker making an approach at a time in order to make the entire approach on his own. But we can have initial identification of these students. The language is irrelevant. The Navy trains Iranians. They train Americans. They train Americans from South, from the North. They train those that speak slowly, and those that speak very quickly. But, it doesn't matter in this particular application because it's indifferent to the language, the student population, male or female, north or south, and/or east and west. We pretrain the system as each student comes into work. As in most isolated word recognition systems, each student has a pattern of his voice for the phrases which are valid for any particular GCA approach. We have found that noisy environments tend to distort and reduce the ability of the system to correctly recognize what was said. This is over all frequencies. The microphone used is the standard supplied with "threshold technologies". It's a close talking, noise cancelling, fixed position, rotate gear type microphone. As I said, we're completely pre-training each

vocabulary item. The trainee says each item three or four times and a reference pattern is created from that. As far as user training is concerned, some people don't want to have to train the speaker, but they want a system that will recognize regardless. Training is inherent in this task. We are in the business to train GCA controller to speak properly, to speak the GCA terminology - the way he will be expected to speak it in the field. Vocabulary is pre-selected and pre-defined. By now we use approximately 45 phrases in the GCA vocabulary, each of the digits, and a number of GCA terminologies. The syntax is orderly and invariant. They should learn the proper way to say RT and the system works very well in training this. We are now looking at a computer consultant-type system. The system overall is intended to grade the controller, teach him how to do the approach, teach him what the GCA approach is all about. Therefore, it serves as a manager of instruction that is being given to the student. We have an idea of what a model controller should do. We have an idea in terms of glide path phrases, course corrections, safety advisories, etc. We use that model to score what he does; however, we expect him to say any of those at any one time. Some students may not know what they're suppose to say and they say the wrong thing. So we can't expect that he will follow the model. The model of the ideal controller is used instead, to score the performance of the trainee. We do make some assumptions about what kinds of errors will be made, this is so that phrases that are "slightly above glide path", versus "slightly below glide path", in an isolated word recognition system - those two phrases would each have a single reference pattern. As you can hear acoustically, there's not much difference between those two phrases. And you will also find that there is not much difference between the patterns for those two phrases. I'm sure you're familiar

87

with the isolated word recognition system, but recall that
"slightly above glide path" would be considered as one word,
if you will, to an isolated word recognition system. Just
as 'zero' would be one word. It stores that entire phrase
as one pattern. Errors then occur. Recognition errors,
while the system curves between "slightly above glide path"
and "slightly below glide path". All those words "above"
and "below". And there's very little in that reference
pattern to distinguish those two phrases. It turns out that
there are approximately 64 bites, out of a 32 x 32 array.
That's very little information. But we have some assump-
tions about when "slightly above glide path" will be said
in the course of training - which I'll get in a minute. The
interaction of the system must be high. The whole point is
to train the user in the proper application of GCA termin-
ology. The system must react to everything that the student
says. Every phrase must react in some way. We have found,
given all these constraints, 95% to 100% recognition, in
terms of scoring, in terms of teaching someone GCA approach.
It's very good recognition. Response time to each input is
approximately a quarter of a second on an over 1200. That
is, I can say, "turn left, heading zero one three, slightly
above glide path, coming down, on glide path", and can re-
cognize those phrases and score the student properly. We
demonstrated this version back in May 1975. We have the
laboratory version currently at NTEC. We've had instructors
from the GCA school in Billington, Tennessee come up for an
evaluation for a week. They brought with them a first-class
air controller who had never made precision approach. She
sat through the system. As I'll describe in a minute, they
were very pleased with the outcome. They had some trouble
with digits and we have corrected that problem. They're
scheduled to be back in December for final evaluation. If
funding goes through as expected we should have a prototype

starting next year. We have a laboratory version now. But want a stand alone prototype which will be placed in the Billington school. Let me just go over the functional design of the system. The first thing that happens is we familiarize the naive trainee with the task, the vocabulary, and the GCART. We give him an introduction as to what the things are about, so that when he's required to make his training tape they're not brand new phrases. Then we collect the voice patterns and make the phrases. Then we go through a prompting of what the trainee is supposed to do. That is, just as in training, an instructor sits behind the student's shoulder and tells him everything to say, we have the system tell the student. The system teaches him each of the phrases that he should say. This introduces him to the task itself. Then we go into a third mode in which the student initiates all the commands that he should in order to make a simulated GCA run. He gives the phrases. We will give him feedback on his performance: we'll fade out the prompts; shift him into the last mode in which we go to a full trainee performance measurement system. We can then diagnose and remediate his progress. We have adaptive training systems that when he makes errors, he either advances to more complex problems or gets pushed back to earlier modes. I am sorry that we don't have time to get into all that; however, there is a paper on the table which explains the notion of automated, individualized, adapted instruction. The speech understanding system is a component of automated, individualized, adapted instruction and it allows the Navy to apply that type of individualized instruction to a number of tasks that so far have been unamenable to objective scoring. The fact that we can objectively evaluate what is being said means that we can objectively score students in their performance. Thank You.

2.4   Dave Carlstrom:

I have a reaction. In that presentation, you each mentioned, at some point, that there was a deversity between what you were doing and speech understanding. It occurred to me though, that, with the exception of co-articulation, continuous speech problem was almost a one to one correspondence. I want to reemphasize this. In the speech understanding case, we also have to build in all these outside mechanisms that you might not associate with speech. Like, what is the particular tasking performed? Because in the show-and-tells next year, in order to prove that they really understood - and didn't just play some trick - they have to follow through and perform some function. And the way I ascertain that the contractor had the right concept model is to have this other domain that, in your case, is the training scenario of that air controller. In their case it has to be a data management system, or something analogous to that - which is outside of the speech thing. So we view "speech understanding", I guess, as it spins off from the word understanding itself. It is doing something useful in the world. That is the only reason you can prove that you are really understood. I see a very close tie, in fact, I wish in some sense that we had known about this and maybe gotten together with you a year or two ago. And worked jointly on some things.

Dr. Goldstein: We didn't mean to imply that there is an inconsistency.

We meant that in our jobs, and that our partic-ular interests as individuals, are not principally concerned with "speech understanding".

Carlstrom: I understand. But, I see a very close match. In other words, if we had picked out (your system)

90
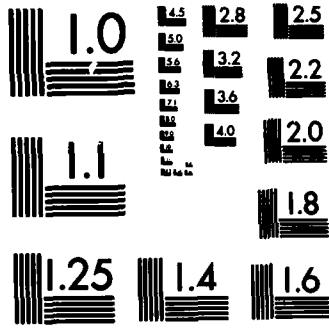
MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

as an application domain one of our systems might look very, very much like yours. It would be very close.

        Dr. Connolly from the FAA would now like to say a few words.

2.4.1   Dr. Connolly: Thanks Dave. I'm Don Connolly from the FAA Experimental Center in Atlantic City. Like a number of other people who have described their project before, I am in the isolated word, speed language, small vocabulary business. Right this instant, out there in the real world, including enroute traffic centers, there are approximately 2000 controllers who before the end of the day will punch about 25 million keys. Automation is probably the good news, supporting more and better and faster infor- mation, but the bad news is that it's semi-automatic. I'm working on a very specific application right now which I believe is probably the most onerous keypunching job in the enroute traffic control system. In this system there is one controller in every position, and there may be 50 of these in a center. These controllers are in the business of main- taining an update of flight progress flight plan information. So, it is automatic that automation can do nothing for you unless you tell us what you're up to. This is the job of the Flight Data Controller; principally. For instance, I found that the language is not as extensive as even I had originally guessed, but what we're working with currently, is the old vocabulary of something under one hundred words. This is divided into not more than about 5 or 6 rather smaller subsets. What I'm trying to do at the present is get a l large volume of very hard data on the recognition reliability. We're using a threshold VIP100 basically, and, after being in business for several months, we are witnessing probably close to a 100,000 increase in the subsets of this language.

91

So far as I have tested, we are working in the neighborhood of better than 95% and, in many cases, better than 99% recognition accuracy. We are in a position to manipulate the language to a limited degree; we can modify the phraseology to improve the recognition, as long as it makes operational sense. We also, of course, will have the user acceptability problem. Air traffic control personnel are used to a very strict speech discipline. On the other hand, there also, like many other specialists, there's a certain amount of inertia against the adoption of new things or in the modification patterns. The isolated speaking message may be a problem that would be insuperable. I see, however, a great deal of hope in the speech understanding business. Speech understanding, practical speech understanding could make a big dent; for instance, in the user acceptability. It could also make a very large dent in the workload. Many a message that must be put into the computer must first have already been said on the air to ground radio to the pilot. And, if we can pick out of that data base relevant materials without repeating it, we'd be really in business. So, I'm an enthusiast in this area. I watch what you do, as everybody in the field does, with great hope and meanwhile I'm working in my little pedestrian corner of this world. Thank you very much.

2.5   The next speaker was John Dixon from the Naval Research Laboratory.

2.5.1   John Dixon: Hi, I'm John Dixon from Naval Research Laboratory. Our group there is interested in artificial intelligence in general and recognition of speech, with the purpose of narrow band speech transmission, in particular. Our hope is to recognize speech on the level of phonemes or something similar to phonemes to transmit these symbols to

get a much lower speech band width. As you know the way
Bell telephone transmits speech digitally, it takes about
60,000 bits per second. We have under advanced develop-
ment now, in a nearly practical state, a system using LPC
coding which transmits at 2400 bits per second. If we can
recognize phonemes we hope to get the bit rate down in the
neighborhood of 100 bits per second. At the present time,
it seems to me that we won't be able to recognize precisely
phonemes but we'll be wiser to recognize a little different
class of speech sounds, which you might call alaphones. In
other words, it may be very tricky to tell the difference
between a 't' and a 'k' but if we have a number of different
classes, develop them so that they are similar to 't's' and
'k's' and we transmit one of these alaphones, then the
listener at the other end, can decide if it's a 't' or 'k'
probably better than the computer could do.

(Editors note: The rest of Mr. Dixon's remarks are not
available due to equipment malfunctions.)

2.6   The remainder of the workshop was devoted to an
interchange of ideas among the participants. Comments made
during this session are reproduced below:

Carlstrom: I really appreciate Commander Wherry coming
down. I am twisting his arm a little bit because he plans
to retire soon and I wanted to get his input before he re-
ceives his orders.
           Copies of the report that the ARPA Group put
together in terms of recommendations for a follow-on pro-
gram are on the table. Since there are not enough for
everybody, if your name appears on the list it means that
you were mailed one. In that case we would appreciate it
if you wouldn't take another one. Other than that, if you

are leaving and want another copy, they will be available
later.

The agenda, as I structured it, is not something
that we rigidly have to adhere to and I am open to sugges-
tions. Future research related to speech recognition in-
terrelationships of word spotting, isolated word recognition,
and speech understanding, are part of the agenda. There
will be some discussion of national efforts already un-
derway. Strategies related to programs of speech and speech
understanding, with messages as how they relate to the
bottom line. That is really why we are here and I think
everybody, regardless of internal disagreements, about "why
speech is" is enthusiastically concerned about speech re-
search or they wouldn't be here. I really think there is
a serious problem about having someone that can champ
this area especially as it pertains to funding. ARPA has
wisely or unwisely, undertaken programs in this area. One
can argue that ARPA has had very adequate funding in this
area for sometime. It looks like that in the future we
are not going to have the kind of funding we had in the
past, and that is why I solicit suggestions, from the floor.
It is possible there are two effects here; one is try to
get a collective effect in funding. (Instead of everybody
doing things in an adhoc way). We need to get more co-
ordination among activities so that important pieces of
technology don't fall through the crack. We want to be
sure that everything that is worth being picked up some-
how gets funded at some level. The other effect is, I think
it is possible to get leverage on management by being able
to cross reference other peoples funding. If there is a
DOD-wide or even government-wide coordinated program it
makes sense. I think it helps me in some instances, to
give more money. That can backfire as they might say
well why fund by ARPA. That can happen but I think in the

present environment we will probably be positive to come in and say, "look here there are hundreds of people who all feel this was and it is very important". So I see that we would gain quite a deal of leverage by building a broad support base with the advancement of this speech technology. I do not think, to be very honest, that ARPA will fund this at the same level we have in the past or in FY 77. I think, however, there is a chance that in fiscal year 78, based on a lot of unpredictable events in the future, one could be optimistic about the future. Again, management could be convinced that an aggressive speech program is important. FY 77 is already on the books and although some flexibility is there it's clear that we do not have the funds that we have had in the past. Also you are probably all going through some of the same things with your organizations, the high-low budget game goes on. So there is some flex- ibility. ARPA could have a little more money than we expect. So I think it is very, very important that a uniform policy develop. I am interested and it is possibly important, to have a follow up meeting of this form with just government people present. I know it is very hard to talk about fund- ing issues with contractors present. I am open to invita- tions. I will come to your facility next time if somebody wants to hold a work shop to talk about these kinds of things. I am willing to pack my bag and go anywhere you want to talk about these matters. Would anybody like to make a suggestion?

Commander Wherry: It seems to me, as others have expressed, we have operated under a handicap for a number of years with people saying we can't do that. I think we need some demonstrations to the fact that we can do that. I think we need to concentrate on the applications that do work. Those that will convince people beyond any shadow of doubt.

Major Carlstrom: Would anybody like to make comment?

Dr. Walker: I would like to make one further extension:to point out that when the speech understanding program first began, it appears that it was explicitly set up as an opposition to isolated word recognition. I don't think however, that that was really the way most involved people in the speech understanding program really felt. It is clear to us now that here certainly is no opposition of that kind at all. There is now much more of a continuum with respect to the kinds of applications which one really needs to address. So that all of the kinds of things that have been talked about at this meeting so far are quite clearly a part of the same technology. One's work is the clearest illustration of all that speech understanding really is. The understanding part of this is completely separated from whether you are talking about continuous speech or isolated word recognition. And, taking the standard word system as I understand it and incorporating techniques for continuous speech will just make any system that much more comfortable with the people using it. So it is clear that is really in the highest kind of spirit of what everything the ARPA program has been asking.

Dr. Christy: There is one particular area in which speech understanding might lead to a tremendous payoff. This is the ability to adjust non-exact word input for the control of a system. I feel that one thing that you can do about accelerating our pace is to make clear what we are "selling" in addition to the specific speech experiments. We might break this out as a special item of a person's ability to handle a positive input and still come out with a reasonable command and control system. The Navy has particular problems along these lines. The sailors that use these systems

are not professionals and are only partially trained.
When they have to deal with their system they are frustra-
ted because before they can come up and cause the system
to do what they are supposed to be doing, they could per-
form the tasks easier the old way. The near misses are
very important and shouldn't be just thrown aside. Try
again. You should try to work with a near miss. This
sounds like what you are saying. One thing about the ori-
ginal comment about showing that we can do it--I have to
be a little bit careful there. Since you are trying to
fund research, it could backfire if you have some success.
I agree some milestones are valuable. If you say you can
do it, then they say, "Why do you want R&D money? Why don't
you go to a protype or production system?"

Dr. Breaux: In that light, the system that we have doesn't
try to "blame" the controllers. Four of the instructors
have only been in training for a while. Except for these
few, progress in the area of research in digetized speech
is limited; however, in continuous speech, we can handle
it now. This is useful. But we still call upon each in-
dividual to have accomplishments in what you call digets.
This is still cumbersome and not too successful.

Dr. Beek: There is a couple of basic problems. We who
are R&D managers have been having trouble with the ARPA
speech understanding system. The information that is com-
ing out has been very poor, except for the last week I
got innundated with all types of reports. This meeting
may help. It did set very poor. We have been unable to
find the applications that can be extrapolated to do some-
thing for us with speech understanding. It is very diffi-
cult. As a matter of fact, we can't do it. We have no

way of justifying why we should do that.

Carlstrom: Well, I guess I should respond. I can sympathize with both of those comments. I just got into this program about a year ago, and I am just now beginning to feel that I understand what is going on. That wasn't because all of the data wasn't available. I am surrounded with it. But, it just takes time to sort it all out. The issue about a common task is something we are concerned about in terms of direct comparisons. I guess I don't know exactly what to say about that. Would anybody else like to comment on that?

Woods: I can see one thing that would relate directly to both. There are a variety of very limited applications of spin-offs that one could tackle with the technology. I think you will find some elaborate natural number grammar that can be utilized. One problem that the ARPA researchers have had in getting that kind of a small bench mark system up is that we have been under the gun with this long list of things we are supposed to try to hatch by a year from now. That sense of real pressure has had its effect. It has not left enough resources to do many of the tasks that might spin-off for application. Some of the evaluation studies that we would like to do take the time to set up are a proper experimental design with the measurements. Check if the information is right. The system has to reach a certain level of capability before we get anywhere. One of the things that you need to somehow achieve definitely I think, beyond the FY76 time period, is to find someway to get those studies funded. You have to have those measurements made to move toward applications.

Don Walker: One further elaboration. In relation to this
common task effort, I think we should realize that it was
uniformly recognized that it would be valuable to have a
common path to prepare a system. On the other hand, what
we saw ourselves in doing the 3 systems, were 3 different
kinds of system designs. Our expectations were that some
of the systems might be able to handle things differently
than others. But we also had, in the context of NASA,
consideration the requirement to take a 20% cut in funding.
The problem became one of being able to do the things we
were expected to do without the review of perturbations
that an extra task would have generated. All of the sys-
tems were originally hoping to have 2 task domains that
they themselves could exercise and get some comparisons,
and it's fallen out that both BBN and SDCS or SRI systems
have to content themselves with 1.

Carlstrom: Yes, that is a good point. We tried about 6
months ago to force the second task to be common. When
we took the budget cut in the program it did get flushed
out. It slipped my mind but that is true. I see a lot
of advantage by having a massive "show and tell." This
would have the advantage of forcing the contractor to
worry about system integration. He has to get all the
pieces to play together, etc. And that is very good.
There is also an advantage for my management. They just
see this thing as one little box. They don't see it as a
complicated set of interrelated processes. They look at
this box the same way they look at a threshold technology
box, or something like that. I need to have a "show-and-
tell" at the end where a guy goes in and walks up and he
expects to talk into the microphone and have something hap-
pen. The same way he would if he went and looked at a

word recognition box. We need to have that kind of a
structure. But at the same time, it deemphasizes the fact
that the subsystem modules themselves are deliverables.
The system of the overall best performance may not be
the one with the best word hypothesizer, etc. And vice
versa, the trick that can occur, the one with the best
overall average performance, may not have, on the average,
the most superior modules. The one that maybe just doesn't
turn over well at all, in terms of total system perfor-
mance at the final demonstration, may have some very, very
sophisticated modules. It's not clear to me that just
having a common task by itself would resolve all of these
ambiguous issues. So there is a real need for people to
critically look at what the internals of these systems are
and how they perform. I agree with Bruno that it is a
very, very complicated process. One argument for a follow-
on program is to just take a year and some resources and
go back and do all of that. Not to develop any of these
things, but just go back and analyze what is there, what
has been accomplished. Since right now the push is just
to get the moon rocks. Now that we have got the moon
rocks let us take some time to study them. No pun inten-
ded on the lunar system.

Bernstein (SDC): I think we are getting into the problem.
We do not yet know how to characterize the pieces, let
alone the system. We are not sure what pieces have meaning.
We are not sure what each of modules are contributing to
an overall system. We are not sure how the modules fit
into the overall ARPA program or how the system relates to
peculiarities of the congressional reviews. We have a
very, very complex issue. We all should have a more com-
mon understanding of perform nce characteristics. We

have to have a better understanding of what the next steps should be. Someone needs to clearly identify that: this is the task; these are the requirements. We must have the knowledge to better characterize where the technology is. Where should we invest the money to move the technology forward? That is one piece that puzzles me. We haven't had time to address this problem. Two years ago we started down the path of working on modules. Now we need to work on where these modules fit into an overall design or overall system. We need to know what any given contribution these modules might make.

Carlstrom: It seems to me that that part of it is a hindsight problem. You can spend a lot of time thinking about how to do right. Then you just say, "Time is slipping by and I've got to dig in and start somewhere." Also, I wanted to point out that a year ago, I think that the only instructions Dr. Licklider gave me when he turned this program over to me was that we had to develop a better metrics and measurements base. He asked me to do that. I think we have been trying to force that issue for a year. All the people have been responding to that. Now CMU had a little bit of an advantage because they had multiple systems, so they could start comparing and contrasting. The others don't have quite as much capability because they have only single systems. But we obviously agree with Bruno. We have been beating that drum. I guess at the same time though, it is the old bit about doing basic research and good science and trying to respond to the pressure to do applications at the same time. We did put out conflicting requirements because we told them that they had to make their domains more DOD-relevant. This was unfortunate. But it is a very real problem in terms

101

of keeping pressure off this research. I think the scientific content can be independent of the DOD relevance of a particular domain for example. If I can do data base retrieval with a reasonable vocabulary I think I can argue that it is not all that costly to retransplant later. But we did have to try to move the domains closer to applications and it's not a bad thing to do but having spent money to do it I think was unfortunate. Having not done it in the first place I think we had to spend some money to do that. We also, at the same time, told the people to try to get common task domains. Things just sort of came apart. We are trying to revector people into too many directions at once. So, in the case of SDC/SRI, for example, they have a status of forces data retrieval thing, which is a very important thing to have in the program from a budget-selling point of view. But it is probably not essential from the standpoint of doing good speech understanding research from a scientific point of view. At any rate I just have to agree with Bruno that that's a problem area and we are all working at it. It probably though will not be resolved to anybody's satisfaction at the end of the program. I think candor is in order.

Reddy: Under the part of the answer to Bruno's question about how to find out what is going on. Our projections are that at the conference next year, most of us are going to be busy working like hell trying to get the systems to work. We are not going to have time to write very elaborate papers, unfortunately. And if some of you need to know what is going on, and it is very important, it seems appropriate to visit for a day. We have all kinds of other visitors coming by. There is no reason why we wouldn't welcome you to come by for a day. Give us a notice; we

can even set up the system to demonstrate. But I don't
think you will have very elaborate reports from at least
some of us for sometime, at least til spring of 1977. It
is just not possible with the few people that we have,
and a number of demands we have on them, to do the research
that has to get done so that the systems are in fact work-
ing by November, 1976.

Hodge: Bruno's comment and Carlstrom's comments related to
the fact that many of us found out that during the first
3 years of this whole program practically none of the pro-
gress reports of any of the programs got into the Defense
Documentation Center. Indeed, the 14 98s were supposed
to describe all of the separate activities that went into
this research. It was to survey all the work that's going
on in DOD. None of the ARPA speech work came out. Nothing
came out of the work in speech research and nothing out of
photography research. This is a problem. I think it
should be corrected. We would like to get that information
into the central place so that we can get it from there.
Or other people can get it.

Carlstrom: It turns out that this is a problem that is
broader than just speech understanding. I was having a
debate yesterday about somebody complaining because some
of the work was not in the Defense Documentation Center,
as it is supposed to be. I think it is supposed to be
locked in as they negotiate the contract. These things
should go to the Defense Documentation Center. Apparent-
ly, because we are not policing the contract, maybe they
are not doing it. But it is a mistake. The contractors
really are hurting themselves by not doing it. Because
the contracts, I am sure, call for it. If they didn't

three years ago I know they do now. But there are still things not going in there apparently. The DDC also is tied into the NTIS. All the ARPA reports are supposed to be in both systems. Contractors are hurting themselves I think by not doing it.

Walker: I just want to say it is my understanding that the blue copies, the copies that go out are copies that should be run through DDC. I have, of course, no way of knowing what happens to them after they leave our office. I have never gotten any response. I don't know if anybody at SRI has receipted for our documents but it is clear that ours have all been sent in that direction. In what I understand to be the problem now, is not knowing what happens to the 22 copies. I am quite puzzled about it. I don't understand.

Bernstein: I agree. One of the essential problems might be the key word index. The contract of the project does necessarily reflect what is going on in research. You know, key word lists do not list "speech understanding."

Carlstrom: Course what happens if I get in that discussion is that I get faulted as a program manager for not making sure that all that stuff gets into the DDC. I shrug my shoulders and walk away. I just don't have the time. I guess it is my responsibility, but I just don't have the time to close that loop. I don't know the answer.

Ira Goldstein (NTEC): I think what has been happening is that a lot of the ARPA reports have been informal that have been exchanged, rather than what we consider formal. I know in our case all of our contractors are conforming

and sending them to DDC when it is a formal report. For instance, if it is a surnote, that is not considered a report so it doesn't get into the depository. It is a matter of definition of what you want to start considering reports.

Carlstrom: How many people know about the surnote system? Bruno, did you know about that? What is the level of understanding? Some of the applications people don't have the background to understand the language in which many of these reports are written. There is a need for some sort of a filtering process. It seems that the RSG group could be the beginning of a form of centrality that could determine what would be of interest to particular locations.

Ned Neuburg (NSA): I am a very strong proponent in trying to get the information around. But I do think that the people are being a little unfair. First place on the surnotes—those are not intended to be published in reports; they are intended to be internal things that are not in finished form and are really for specialists in general. Now I guess reports have been coming out as surnotes but that was not the original intention. I think a lot of the information that had been coming out of there, out to the project, has been coming out in perfectly reasonable ways. For example, the proceedings of the conference that Rog Reddy held, at Carnegie Mellon, includes a lot of papers by people in the ARPA group. They are just at the level at which the work is going on. They are scientific reports and it is true that a manager can't read those reports and decide how the project is going. But, on the other hand, that is the nature of the project at this time. The application has not yet been tried really. One can't say it is going to succeed and therefore you should pay

attention. You can only say we are working on certain LTC
techniques. We have a grammar that will do this etc. etc.

Carlstrom: At the same time though, I would like to point
out that I agree with everything Ned said except the sur-
notes. I don't think there is any major objection if peo-
ple doing similar research, or trying to build similar
cases for research etc., had access to them. As long as
it was recognized by all that they were unofficial and are
internal documents. By the way, a lot of official things
that are released do end up as surnotes. That practice is
a good one. I have never gone back to my own contractors
and said so, but I would say so now and I sort of like
that practice. A lot of the technical reports that have
been submitted have also been placed in the surnote inven-
tory and I think that it is a good idea.

There is something else that seems worth track-
ing. That is, the problem of marketing what we are doing.
The translation aspect and somehow it seemed worth trying
to hang on to. I don't know where one goes with that but
maybe it would be important to have a symposium - you know,
classical music for people that don't like classical music.
I have that problem too and I can really empathize. This
last year has been a very grueling year for me. Sorting
a lot of these things out. I knew quite a bit about sig-
nal processing but I knew practically nothing about speech
etc. All these different things. It has been a fun year
but quite a sabbatical for me. I think I can really under-
stand the problem of other people who are trying to work
requirements or work against ROCs, etc. They don't know
if they are being Buck Rogers when they estimate that cer-
tain things should be done by such and such a time. Maybe
there is some way to have another meeting to try to ad-

dress that problem more directly. I would be open to communication on that. If anybody wanted to try to make suggestions to me about how to cope with that problem, I would be open to suggestions. It seems to me that we are talking mostly about the sur program. I wanted to try and get off of that for a bit. One of the things that the sur program did not try to take on was on the speaker variability problem although it did allow for a slight amount of normalization. You know they talked about slice tuning. They also specifically excluded training, hard training on single speaker. They wanted to have not a strong dialect, like one might have from Brooklyn, or the Bronx, or Alabama, or something. But, they said a general American dialect and in the demonstrations I have been in that sort of work. There were 10 people in the room, and the thing seem > to work reasonably well, out of the 10 people! Although the ARPA's systems are some sort of a norm. They didn't talk about normalizing, adapting completely to individual speakers; on the other hand, they didn't expect to work with just one speaker they were finally tuned to. I know that most of the people in isolated word recognition at one time or another have talked a lot about the speaker variability problem. Maybe some of the people- not ARPA people-would like to talk some more about that. What do you see as the next step? If you were going to go out and fund research tomorrow and put together an experiment tomorrow, to advance those problems, how would you do it? You know, if you were going to put dollars together in a budget to go do something there. Bruno, do you have any thoughts on that?

Bruno: You mean in terms of speech understanding? Or what?

Carlstrom: Just in general.

Bruno Beek: There are efforts underway now. But we are
looking at the problem of speaker variability. We are
trying to compensate for that. Both Ned and I are funding
small efforts. We have been looking at the problem in
general and we have also been looking at the problem for
specifics. For example, in the speaker verification prob-
lem, I mentioned that we need either to keep key-board en-
try or some sort of ASCII. What we would like is to do
code recognition, like 4 to 6 digit code recognition.
Doddington, at Texas Instruments, is putting the system
together for us which is a speaker doing continuous strings
of digits. Except, there is a cop out on it. Because
first of all we use error correcting codes and also cer-
tain digit combinations. We don't have to worry about
the segregation problem. We have been looking into that
problem. We think outsiders looking at the ARPA problem
seem to be very cavalier about this. We don't know how
many speakers you are going to test. What kind of tuning?
How is the system working right now? Can we actually take
some of the work that you have been doing and relate it
to our problem area? We don't have that gut feeling. I
attended these meetings and we talk about limited experi-
ments and overall strategy, chronologically rules, and what
have you. But we don't have that sort of hard information
actually coming out. That is the type of thing that I
think bothers most of the people who are working with spe-
cific applications. We are addressing ourselves to their
problem. We will continue. Because, for example, in key
word recognition, even language identification, speaker
variability problems is programmable. If we could solve
that, we would really have a good handle on handling the

word spotting problem.

Goldstein (NTEC): In our problems with speakers, different speakers, we've brought them through a few radar approaches. This is to get them ready for, and used to the GCA task. Essentially, we give them homework on what they will be doing. Then we begin to stage the actual experiments until they get used to what they are doing. We found in both studies that those individuals who went through the pre-training, didn't have the problem that the curves predicted. It became an everyday thing! It became a boring task. This approach has been incorporated into our system.

Carlstrom: Of course, the difference there is cooperative versus uncooperative speakers. I am not sure though that in the uncooperative speaker case, that it is an open loop situation. From the terms in your point of view, you don't feed back to him to set him into any pattern. But the system that he is operating in is closed loop, so he is really not uncooperative. He is cooperative with something. You should be able to get some leverage on that if we got smart about how to do that.

Neuburg: I think the problem really is 3 kinds of speakers and not two kinds. There is a cooperative speaker, there is a not-uncooperative speaker. That is a person completely unknown to you but he is trying to communicate with somebody else. And the, in between, there is the person who is going to use a friendly system and all that, but you don't want to have to train him. There is really a third case still. Because he is cooperating with the machine. The only thing he is not cooperating with is a

training algorithm if you like. He is not going to sit down and say each phrase five times. I think he is the person that the ARPA effort is really aimed at. I think in the back of everybody's mind, really when they wrote the specifications, was that we should be able to let the man in the street come up and operate this machine as soon as he knows what it is trying to do.

Carlstrom: I think that is fair. I have written that many times in the internal paper work. The ARPA system is really trying to replace typed input for the lay operator of the machine, a non computer scientist type. But I think it is also true that whether or not we go directly after some of these other problems, we are right in the middle of them. Possibly we have contributions to make if we can ever find the right way to talk about them so people like Bruno can track. See what we are doing and what we have got.

Graham Gross (LTS): I think that in some of these applications, there is more continuous interaction by the user and his assistant, that there is no need to compromise on the system parameters. I think that some of the applications that people are talking about meet the requirements to some extent. But, you could conceivably "tune" the operator a little bit. If he speaks a word that is incorrectly interpreted by the system, it should be displayed immediately as if it is a misinterpreted word. Then the subject has an opportunity to say it again. In this way, you are training the subject. There is a bio-cybernetics feed-back sort of thing. You could have the system with a constant echo. It may turn out to be too annoying.

Goldstein: This is where our system works-during that

training phase. The whole point is trying to impart feed-back. As he goes along and makes a mistake the system briefs him and tells him what he should have said. So if the aircraft crashes, he gets a feed-back on how he said things.

Carlstrom: I guess I really believe that. The problem is that with the lay user trying to use the computer, he doesn't get helpful feedback. That is why he feels he wants to kick in the front of the terminal. He has to go to a library, that has wrong documentation anyway, and then try to figure out why the darn terminal doesn't do what he wants it to do. If he got the right kind of feedback he is willing to learn. He just wants interactive good feed-back. The other thing is that if he is typing it isn't clear that any feedback in the world can make your fingers work right when you just don't have the knack. So, there is feedback on two levels. One, where it looks like most people can modify their speech to elicit the right response. I know that is true because of flying an airplane myself, and talking in on radios. When one first gets in that en-vironment, he swears up and down that he never is going to be able to cope with it, and it doesn't take very long be-fore he does. You can understand what is going on and talk back through the environment. Also if you go into supply inventories, or logistics scenarios, the same thing applies. You first go in there, you swear they are talking Greek. You are in there working in that environment, and sometimes it takes only a couple of days, and you are right in the middle of it throwing the jargon around with the rest of them. I agree completely with what you are saying. Even for lay operators you can expect them to train up quite a bit. Are there any other comments in that area?

Marbury: I am here today because I just had a recent bring
up from all the DNA information centers and looked very
carefully through the last 10 days for any information on
recent reports. I found only a few scattered around. We
probably have one of the most complete scientific techni-
cal libraries in the Metropolitan Washington area. And
yet we have very little information on speech understand-
ing and that is the reason I am here today. We have some
systems interest in your field of effort but it is more of
an interface interest that we have rather than a total sys-
tem interest. My own immediate problems are in the field
of nuclear security and we have little "carrots" in the
way of 6.2 and 6.3 money in FY 77 and 78. But, you have
got to show me a direct application in nuclear facilities,
weapons, or especially nuclear materials. There is an in-
terest in what you are doing but basically for an inter-
face device. It could be a dedicated speaker as far as
that is concerned and have a security interest. We have
a dedicated site facilities that is a complete system in
itself. Many of you from the Services are aware of this.
We have the requirement for the interface device. There
are 3 or 4 of us here for that interest. Now to the ex-
tent that we would be talking about the interface, and to
the extent that this group could find some ways to talk
in that frame of reference rather than in the total sys-
tem, we could possibly join with you in spending 6.2 and
6.3 money. But as far as I have any direct relationship,
or my agency having any direct need for a total word re-
cognition system-as we have discussed today-we really try
to keep our words off the line, rather than put them on-
line. So, we are looking at that other end of the game.
We could, to the extent that this group would be interest-
ed, sponsor a meeting such as this. We could get with

ARPA and put some of our nickels with their nickels. But this would only be in support of nuclear-effects research.

Carlstrom: We have talked before about the related scenarios. Essentially computers are being used for security of nuclear devices. Computers will depend, more and more, on knowing the risk level of storage facilities. What DNA is interested in is having the capabilities of guards, and so on, being able to input facts about a situation quickly into the system in a natural way. If a guard sees a gate open, he wants to talk into a Walkie-Talkie, and say, "Gate 35 is open." The computer will then say, "Well, that's OK." You don't want the guard to have to type in. This is the general idea of Marbury's interest.

Marbury: I'd like to say further that none of the contractors are likely to hear direct from our agency. The contact would probably be made through the Service Laboratories or DARPA, or existing contractors with the agency. Basically, we won't take the lead. This is an ARPA lead. We don't get in conflict with them. I'm sure Dave and I and Dave's boss and my boss can get together in time to make the FY 77 time frame. We can't make FY 76, but in FY 77/78 we can give you some encouragement.

Christy: I have a comment on the extension of the technology. At the present time ARPA, and others, are reviewing the results in speech understanding systems. I wonder about looking to the future after we have the "front-end"- the speech understanding part. It seems to me that the applications part is the part we can solve. The front-end can be useful. But future research should address what "has to be!" What application has to be; what specifica-

tions has to be; what criteria has to be? The front-end should be the tool to help the pragmatic "what has to be."

Carlstrom: That's a good point. One of the strengths of speech understanding is that it tries to integrate natural language kinds of things with speech understanding portions. It tries to view those in some synergistic way. However, there are other ways to cut the cake. With the problem of funding over the next few years, it looks like natural language will be fundable in its own right. In the past, it had to be accomplished under the guise of speech understanding. I see a problem in doing more in the natural language area, in that the rules change, there is no punctuation, etc. Nevertheless, funding is bound to find its way to natural language endeavors. It is possible to ease some of the budget strain by funding a lot of that type of work in a natural language setting. I think the contractors already sense this. The disadvantage of this is there is not a one for one correspondence between these two domains and the interface between the lower-level processing and the higher-level processing is not a clean one. What this may mean in ARPA is that our funding should concentrate on the speech side of the thing and we should carry the natural language efforts under a natural language banner-in the budget sense. But I believe in the whole spectrum of technologies that should be advanced here.

Christy: My problem is deeper than that. When you decide on an application, how do you extract from the application task-the semantics-to help structure the speech understanding or the natural language program to fit the problem?

Carlstrom: I agree with that. I think this can be one of

the contributions that the speech understanding work can make.

Marbury: I can see a bunch of "carrots" here. If you can come in multi-lingual, and have double language inputs, and still give me the printed output, this is desirable. I don't know why you can't do it in the FY 78 time frame.

Carlstrom: There is a very interesting trade-off going on there. English translation became a dirty word a few years ago. This is no longer the case. A lot of natural language people want to do language translation.

Marbury: We have just put out a new document-world wide- that requires all signs in bi-lingual minimal.

Dave Walker: We certainly have had a great deal of experience in that area at SRI, both in the speech understanding system, where the term "great" does not apply. Since we are building only two basic systems in our artificial intelligence center. That's been one of the major concerns of artificial intelligence over the years. Something could be said about a good number of other places in the country-Stanford, MIT, Carnegie Mellon. So I think there is coming to be more and more unity of purpose. I hesitate to call it a consensus; it's probably not that far along. But I think that the kinds of experience available now allow us to feel much more comfortable about going into new domains, and be able to see how we can adapt the technology. We have full representation for the manipulation of concept structures in these different areas. As I say, in particular, our work addresses both data management and what we call a computer-based consultant problem, where you have

a very specific maintenance model with electro-mechanical
equipment. We found that our overall semantic modeling
and discourse structure modeling works very well in these
two very different kinds of domains. So, these are the
things we have been able to do under the available resour-
ces. We would like to explore systematically. It is some-
thing we talked about in ARPA speech understanding program.
One of the major concerns is how would we in fact extend
out work into new task domins easily. It is clearly some-
thing we see as a major requirement.

Marbury: There is one other point before you get away.
I am not speaking in terms of bilingual intelligence func-
tion or any board activity but I am speaking strictly in
an overt security mode where the bilingual vocabularly is
something less than a few thousand words. In fact, merely
200 words. And as to what language it is spoken, I am not
concerned with that. In other words, in a physical pro-
tection mode, in NATO, or US forces, or anywhere else.
We are speaking in terms of host-nation language and guest-
nation language. That sort of thing. So, we are really
only speaking of day-to-day interchange languages and
"what is the status on the reactor site" or "a weapon site"
or something like that. This is open language and we just
want the status reported. Sometimes it might be in English
and sometime in broken English. But the words would mean
the same. It is just the question of whether or not you
can recognize them and give us a basic package. One that
will go into a monitoring or computer-monitoring base. If
you can, we could talk about that and I think we can get
some funding in that area.

Carlstrom: If the person involved is a foreign national,

employed by the U.S. Government, you might be better off to let him speak in his natural language than to try to use another tongue.

Marbury: We are only saying that the vocabulary in a physical security mode is not a great problem technically.

Carlstrom: If anybody has any thought they would like to put on the table now, go ahead. Now, one of the things I thought might be worth tracking would be to get some of the ARPA people to talk about what they view as the variability problem. I know from having reviewed a couple of proposals in the last month or so that they are talking about this. My view of the proposals is that we would like to get into those things. A lot of luxuries are there that probably won't go. But, I am glad to get their ideas down on paper. Because, things do occur to them. Perhaps some of the ARPA people would want to comment on any ideas they have had about speaker normalization, and speaker variability. Just what kinds of parameters are you talking about normalizing? Would anybody like to address that?

Ritea: There is a whole list of issues. I don't think we should just discuss speaker variability, I think we would like to talk about environmental noise, dialects, and the number of other features that should really take into account the application systems. Do this within the present framework of the goals of a speech development system under very antiseptic conditions, if you will. We haven't really had as part of the program, the time, nor the resources, to investigate a lot of these issues. Rog Reddy has investigated the telephone channel problem to a limited degree. But I think I would like to see the discus-

sion centered on a whole host of issues relative to an application system. Which ones are really foremost? We would hope to implement a system like this in the next several years. Is it going to be a dialect? Is it going to be environmental noise? Which one of these areas can be controlled? Which ones can't. Which ones can be solved by using the noise cancelling microphone, for example. Which ones can be solved by controlling the amount of background.

Carlstrom: Would anyone like to comment on that?

Reddy: I can make a few comments on what we can expect. For example, if you go to some of these telephone lines. What happens is, what we all know is, that some of the tricatives in the computer get lost. We also lose at the low end the nasals which are much lower amplitude and they look like voice stops. So on voice tricatives, and particularly on low amplitude voice, nasal and voice stops will be confused with each other. So what has happened in any system that does symbolic matching is that it can't do two kinds of things and expect it to match some electrical entry. It shouldn't be too picky about which one it picks. Right now I guess most of our systems either expect one or the other and they go ahead and do that. So, one of the noise normalizations comes up to the top word matching level. It does not come at the level of normalization. The same thing happens when you correct it with noise. That if certain kinds of sound, especially low amplitude sounds, will all get corrected and one can't distinguish between them. Many of these voice stops and modules look like voice tricatives perhaps. And again you need to have techniques at the word matching level to normalize for this.

118

None of our systems have it now but we think we know how
to do it. Again, none of us will be able to get to it un-
til after the end of next year. First, just because we
know how to do it, it doesn't mean that we may actually get
it into our system. So much for the noise. The other is-
sue on the speaker normalization. It is our experience
that the dialect variations are causing an order of magni-
tude more difficulty than the vocal track variations. I
don't know if that makes sense to others. But, that has
been our experience. We did have to work much harder to
take care of inter-speaker differences. Because it pro-
nounces slightly different and a little bit sloppier.
This is rather than the vocal track shape normalization
which has been very much the work just done in the past.
There is no substitute for studying the dialect and going
and studying it and understanding exactly how to do the in-
terchange and then represent it.

Carlstrom: That seems to me what I have to do to under-
stand dialects.

Beek: From all the systems that I have seen to date, re-
lated to noise, the isolated word system, speaker verbs,
verbal communication systems, all have got troubles when
you start talking about noise and band-width constraints.
They just don't work very well. In some of the tests that
we run there is quite a bit of controversy about that. We
still argue about this quite a bit. But from the test
we've run at Rome, they just don't work very well. We don't
know why they don't work well, or, as far as speaker nor-
malization is concerned, in quite a few of the tests that
we have run we also have found two types of speakers. One
that we call "sheep" and one a naive type of speaker. The

"sheep" are beautiful. They are usually the people that
work on speech recognition (lots of laughter.) But we
generally experiment also with these naive people. They
give you all sorts of troubles, and our performance evalua-
tions here are consistent. Errors are consistent among a
certain set of people. So you know it's a real problem.
Especially if you are starting to talk about something
that people can really use in the real world. That is why
I am a little concerned when you say that ARPA is going to
test many speakers, many cooperative speakers. Are they
cooperative speakers? If the system doesn't work well for
them, we won't worry about that! Or, will we know how well
the system will work for many? What do we mean by many?
In some word recognitions that I have seen many means more
one. Or one as a matter of fact.

Reddy: I think the expectation is between 5 and 10. We
do notice that it isn't going to be a 100 for example.

Beek: One other question--Are you going to give a test
plan? Are we going to have a test plan at a time to find
out how these various systems are going to be checked out?
How many speakers? How many sectors?

Reddy: We talked about it at one point. But, he didn't
do anything about it. He said he had some ideas about what
it ought to be, then he didn't tell us.

Carlstrom: It seems to me that in some sense the ARPA sur
systems have defined their own game and they will either do
well in that particular scenario they have defined for
themselves. And, be criticized because it is a trivial
scenario. Or they will pick a scenario that people will

deem incredibly difficult and will do poorly and will be
criticized for that. Or, if they are lucky they will have
picked a difficult scenario that is incredibly difficult,
and looks like it has some utility in the realy world, and
they will do reasonably well at it. So it seems to me that
the testing plan cannot be external from what people have
already done. It seems to me that the only thing you can
do, in a systematic scientific way, is to do it in terms
they claim they have built. And then if you feel that
they haven't built something meaningful, that can be part
of the criticism. I am not a disinterested party at this
point, but I think I could go in with my level of knowledge
of speech systems and design a test plan. I shudder at the
thought that I might be asked to do that i the next 12
months. Mainly not because it wouldn't be a fun thing to
do but I don't know where I would find the time to do it.
But I think it is a fair question.

Walker: I would like to make one observation. I have
made this observation several times about the lack of spe-
cificity. I think one thing you should realize is that
the system builders were sort of confronted with a 19 var-
iable design. A lot of uncertainty exists about what
things were going to work. And at what level. I think
we are all focusing on the test plan. We will be making
commitments. Commitments that, if we had been smart enough,
we would have been able to work out four years ago. I
think we would probably not have done a very good job if
we tried to do that. There has been in process, in other
words, some sort of growing, and becoming, for our system
efforts. We have had a fair amount of consistency over the
life of the program. I think we do have to have test plans.
In those test plans there is certainly going to be more

explicit commitment on the lines that you are suggesting.

Woods: I think one of the things you have to realize about
that list of objectives, in a 5 year program was that it
was an attempt to project what somebody thought might be
achievable in 5 years. And by looking 5 years ahead, to
try to be much more specific, as to what you mean by many
speakers, and exactly what you mean by the general Ameri-
can dialect. There isn't any general American dialect.
We all knew then that there is no general American dialect.
There was, more or less, the understanding of the notion
of many speakers-not spelled out-which ones, or a random
set was indeed there are sheep and goats! And we are con-
tent to deal with sheep in the five year program. There
are a variety of applications for speech understanding
system where you got control of the person you are going
to employ as a communicative machine. You want to talk to
it whenever and you want to have an aptitude test. I as-
sume is a detector that works pretty well. That would be
appropriate to screen the people that you're thinking of
employing as controllers of the machine using speech. We
felt that was indeed one of the dimensions you could make
an attractive problem out of. To say we will deal with
the multiple speaker problem, in the sense that we don't
want a system that you have to train specifically for one
speaker. But, we won't go so far as to commit ourselves
to take any old speaker you bring in off the street. We
want to reserve the right to rule out a speaker that's got
extra performance or holes in his former track. Or any of
the other strange things that we get. The voice breaks up
in the vocal project. Is there a difference in primarily
male speech or female speech? All of these things, in
that outline, are targets we are shooting for. Many of

the projects are going beyond a lot of those targets. We
are looking at many dimensions in noise, and in speech.
We are not trying to commit ourselves forever. We are con-
cerned about the dialect problem. We don't expect to solve
it. That list of things is really a list we are aiming
for. We have a lot of things set up for 5 years from now
in anticipation of events. We have set up very meticulous
descriptions of voice testing methods.

Beek: That is sort of the information I need to know so
I can make projections to the R&D manager. He needs to
know what we can do in the near future and in the long
term. I can't work only with machine type problems. I
am not going to be able to force this into a sheep and
goat situation.

Woods: So we have to add that into the noise. We've just
read reports so we know what is going on at the module le-
vel but we really want to know what the whole system is
going to look like. I guess you don't know yet either.
So for us to project what the future needs are is very dif-
ficult. This is as difficult fo us as it is for you to
project what you are going to be able to do by November.
I would like to reiterate however what Rog raised earlier.
The best mechanism I know of to get that feeling is for you
to visit the site, visit all the sites. Get to know the
details in more depth. There are so many dimensions of
these very different tasks that we can't study them all.
But if you lay down the specifications for particular tasks,
I think we are developing pretty good intuition as to which
things will be the troublesome things; which things will
be difficult; and in fact we could, within the year that
we have left in the program, definitely plan to do some

123

experimentation.

Carlstrom: There seem to be targets of opportunity that
pop up. It is like the system builder who would not have
been so bold as to say that, "I am going to do this" but
once he gets into the problem he sees that he has the pa-
rameters right in front of him. So, he just does it on
the sly and keeps on going. It depends on what you mean
by speaker variability, but I know that some of the people
are looking at speaker rate and vocal track length. Things
like this. They don't promise to solve these things but
they are going to work on them because they are right there
in front of them and they are fairly easy to do. I think
that is why they made such broad statements in the begin-
ning because they felt there would be things like that along
the way. That is why they said many speakers instead of a
single speaker because they wanted to leave the door open
to go after some of those kinds of issues. Yet, on the oth-
er hand, they didn't want to promise to emphasize Bill's
remark. It seems to me that the task of finding the appli-
cation that emerging technology can provide is a two way
thing. We should tell you as much as we can about what we
are doing and what we expect to accomplish. But the more
you tell us about applications and their requirements, the
better the style can go on. I think in particular of Com-
mander Wherry's example, and Don and I probably have the
same reaction to that. I see that as an example of a per-
son who attacked the problem without knowing about the ARPA
speech program. He used existing speech technology. He
really has speech understanding system, which now clearly
can benefit from connective speech input and a different
kind of training procedure. He needs new technology from
what he was using. Finding about his application helps us

answer the question of what can we do for the real problem.
This can be beneficial. So, all I am trying to say is that
I think it can work both ways.

Neuburg: Just as sort of a technical point, there is ano-
ther large speech effort being funded by DOD and it has a
certain interaction with this project. In fact, some of
the interaction is sitting in this room. It is the big
speech compression effort. There, they are very worried
about being able to use many speakers. Because I have to,
and also using a telephone, because they have to. I think
that in a sense it takes a little of the burden off this
group because somebody else really does have to worry a-
bout that. In fact, it turns out that many of the para-
meters they are extracting are the same as the parameters
beign extracted here. So, it is almost as if you had ano-
ther group that is working in parallel, and looking at the
same problem. I know it is not the same problem as worry-
ing about the mobile speaker and poor channel problems.

Carlstrom: I am glad Ned brought that up. I meant to make
a statement at the beginning about both ARPA's speech pro-
grams. I didn't do that. I am program manager of the
Speech Understanding Program and Dr. Con, in the office,
is manager of Speech Compression Work. We sort of cover
for each other. Because, in the front end, parts of the
system overlap. We borrow from each other; from a coopera-
tive stanpoint. This isn't true across the board and in
all the applications. But in some of the applications the
problems that people have run into will benefit more from
the Speech Compression Work than perhaps from the Speech
Understanding Work, per se. I have spent a lot of time
dwelling on that but I just wanted to point out that there

is another system program in the office that is part of the DOD speech consortium and a lot of you know about it. A lot of the same specialist contractors, work in pieces of both programs. A lot of the exact same algorithms are in both programs.

Reddy: I have a question. There was an issue raised about noise. I want to find out how much other people know about them because we have been having a great debate about whether they are any good at all. Because when you look at some of the characteristics, the amount of noise cancellation that it gives is about 3 B, between 1 K and 5 K. 3DB is like half a bit from the sample. Is it really worthwhile getting a noise cancelling microphone? If anybody has any comments it will be appreciated very much.

Hodge: Some new techniques in this area are being investigated, partly by the U.S. Army Electronics Command and partly by the Army Air Medical Research Lab at Ft. Rucker. I am not too familiar with their use. They involve the use of condensor microphones. The problem is that the Army is getting an integrated communications system and the typical condensor microphone has to be polarized. Therefore the microphones are not directly or immediately, compatible with their acostic communications systems. So there are some problems in getting these into the system. But, apparently they have been able to succeed, and achieve great success in suppressing the noise in helicopters. In the helicopter environment, the talk is usually better. I can get you some names, but I can't give them to you off the top of my head.

Christy: Your figures are rather amazing because generally

126

we have 12 DB, not 3 DB.

Reddy: There is a peculiarity about that, between 100 hertz to about 500 hertz is like about 12 DB. Then comes 500 hertz to one kilohertz, it goes down more or less to 3 DB and it stays there. So most of our interest is between the 500 hertz to 5000 hertz.

Hodge: Most of our commo people are not worried about that because we don't transmit in that frequency range. We have got low frequency noise problems to start with.

Reddy: If you have a low frequency noise then it does help. We are working on noisy environment but they are more like 60 DB or 70 DB. But not necessarily very high noise environment. And so the question is, is it at all worth while going to such noise cancelling mikes?

Christy: Many of the Navy's systems use noise cancelling microphones. So, if one is considering going toward that type of application they are going to be there so it is not a matter of choice.

Hodge: You touch a nerve there; my specialty is acoustics. We have found in two different studies--one involving aircraft and the other is armored vehicles--the primary source of hearing loss to personnel is the high level of noise in the communications system which is being picked up by the noise scanner.

Mundie: The Air Force solution to the problem is to put the microphone in a protective device like you wear over your ears when they are working aroung the jet engines.

The microphone is sitting in that sort of a noise eliminating situation. Not noise cancelling. You protect the microphone just like you protect the ears.

Carlstrom: I think the issue is the guy in the cockpit itself. Everytime he hits the mike switch all that stuff blasts in. If you put him on the oxygen mask all the time that doesn't work. I know that phenomena because one aircraft position I flew we had hot mikes. You left the hot mike on because of the two guys working together. It was easier to coordinate if you didn't have to reach for the foot switch or the mike switch. But, you would get the fatigue from that because you woul hear all that cockpit noise through the headset. It would be for 20 to 30 minutes and it would start driving you bananas. So you would flip it off and then you go on the other mode for a while. That would then irritate you, and you would go back. And so, back and forth, on a 10 hour mission. You spend alternate 20 minute cycles, with hot mike on and hot mike off. Does anybody have any other issues? I think I have missed some hands at various time and I don't think we should hesitate to go back and pick up anything left laying idle. Does anybody have anything they would like to surface, any new discussions they want to raise?

Neuburg: A technical issue. I guess there are really a number of reasons, but some of them very immediate and some sort of long term. Why are we all sitting in this room at this moment? I guess the immediate reason is that ARPA has identified a crises or whatever it is. A budget crunch at the same time. But, there are also some longer term reasons. There are very good reasons for all the R&D managers who are sitting here to be sitting here. They should

be exchanging information and they should be listening to the people who are doing the things and discovering what sort of technology is available and that is becoming available. When this workshop was first being discussed I was very dubious. But I now realize this is a rare event and an event that will take place again in some form. I would sure like to hear if people have ideas of how to make this a more or less continuing thing. I know that the R&D managers need it. I also know that the contractors here need it. They like to know what sort of opportunities there are. I have never heard of your application. How can this be carried on?

Bernstein: I think that I would like to hear a bit more from some of the people who haven't said a word. And there may be at least a half a dozen over here who may be able to tell us what their perception, dream, desire, or what have you, is for a speech understanding capability. What are their needs in their immediate and long term goals. I'd like to get a more comfortable feeling that we are headed down at least one right path. I have heard a reasonable amount of criticism about the fact that ARPA is working in a rather sterile environment. Supposedly we set our own boundary conditions. We set our own objectives and goals. We use our own strategies. I for one have some difficulty with that. I am sure other people have difficulty with that. This is one of the few time we have a large respectable set of people with requirements. But the problem is that I am not hearing much of a discussion on requirements, as opposed to a discussion about technology. I think I would like to see a little bit of a shift in direction and hear more about what people think they want out of the research in terms of applicable technology.

What kind of boundary conditions is a better fit? What is
going to be useful by somebody in their set of perceived
problems?

NTEC: People sit down and specify exactly what they mean
along these dimensions. Contractors can certainly consider
that perhaps. Along the lines of what you are talking a-
bout, if these kinds of meetings occur more often then the
contractors who don't have time to write reports, and
don't have time to get into the information system to dis-
cuss what is going on can relate here. They can also give
us a run down on what is happening.

Neuburg: In spite of the fact that the meeting got changed
twice on very short notice a lot of people showed up. It
isn't all that hard to get here it turns out.

Medress: I would like to amplify your point if I can.
Those of you who have read the following program plan, and
those of who will, will see that we tried to argue that one
of the reasons there should be an ARPA program, is to take
advantage of the technology that was developed in the first
5 years. After the first five years, then get the present
program feasibility demonstration. We thought pretty con-
fident that good technology was coming out of that and that
technology could solve some real problems. In trying to
make that argument we didn't have a good handle on what
the real problems were. We had to do a lot of guessing
and we had to do a lot of sketching, etc. So I really
think if that philosophy is right we really need to get
more specific information about the applications, and that
works both ways. The more information we have, the more
we can vector that kind of development. You know, in a

130

in a useful way.

Carlstrom: By the way, as a footnote there, I welcome cri-
ticism of that report. There were two arguments. One was
we should have had this meeting before that report was gen-
erated. That was the intent for a while, and we just could
not squeeze it in. On the other hand, maybe it is better
to let that report be a catalyst and let people find fault
with it and submit criticisms. I would welcome letters
from RADC, from Aberdeen, or anywhere for people who think
that report doesn't quite state the problems just right.
Or that it doesn't cover some requirement for the out-years.
I welcome that kind of a dialogue.

Mundie: I would like to return to the opening remark by
Commander Wherrey. He said that there is no substitute for
success. I believe he said virtually that quote, and I
think that is very essential to the program. This has
been an ARPA multi-million dollar program. Unless some-
thing comes out of this very quickly, and something shows
up in the Dept. of Defense to make use of this, everybody
here is in trouble. At least in terms of financing things.
The environment now is that you have got to tie your re-
search to something that gets very quickly into applica-
tion. So I would like to open the question, and then fol-
low on with the suggestion, as to where the Dept. of De-
fense could make optimum and quick use of the technology
that is going to be available from this program? I think,
to quote from this follow-on report, it is obvious that if
you have to add a terminal to a computer, where you could
speak to it, and it spoke to you, and given the choice,
people would choose that as their terminal to communicate
with the computer. I think voice communication is the

131

channel of choice. I am convinced that it is just a matter of time until this happens. I think it can be solved. I would hate to see a technology flounder as near to success as it is now. Just because we can't direct our attention to one successful demonstration and put this into the field for use. I would submit that a possible choice for this is the one already launched by the Navy, the teaching situation. This can be highly structured. In terms of the language that is used, you can be highly selective. You're working with only one student at a time. But, teaching situations are present in all three of the services where training is a major undertaking. And, in more places, the training is being done by computer instruction. There is an interaction of the student with the computer. I think to put voice input and output into that situation is a practical application. It allows us to work with large systems. We don't have to worry about compressing it down and putting it into cockpit and that sort of thing. It is highly structured. It has many advantages for making a success out of speech recognition. I think speech synthesis is really no problem any more. We would agree with that. Now, you can take a phonetic strain from a computer and produce understandable speech from this. So I would submit this as a problem that is a reasonable task. It includes speech understanding, speech recognition and speech synthesis. It is a realizable task in from 1 to 2 years if we devote our resources to that particular problem. Then we would have a successful demonstration where we can demonstrate that man can converse with the computer.

NTEC: Some of the continuous components could be incorporated consistently. There is a demonstration system now. There have been a number of people that have been down to

see it. It is an interactive kind of thing. It all depends on what dimensions you want to demonstrate.

Carlstrom: I agree with all of this. You have to be very, very careful. Some of my managers, who will go unnamed, visited places when I first started talking about speech understanding system. They said, "Well, the Post Office has been doing that for years." When they see that success, in their terms, they just want to eliminate all funding for any R&D in this area. So you have to be very careful to make sure you educate the fellow ahead of time before you brag about the success. You make sure he understands how to interpret that success. Be sure he interprets it as a milestone, or as a progress along the way. But not as all wrapped up and done. Or, that it is on the assembly line and we are ready to turn them out.

Beek: The point that I tried to make is that we are building the system. The Army has a training system. They talked about it at the Juarez meeting. There, they described voice input and voice response. The Navy has one. There are a number of other systems that are using it. We have built speaker verification systems. We are going to use voice input and output for the photographers. Because of that work, it has led to the photo interpreters saying, "Hey, what can you do for us?" They work from stereo plotters. They have to make measurements from photographs. So, it is a problem of going beyond the digits now and talking about a structured vocabulary. We can do this type of thing with isolated word recognition. We think we would rather like to do it with connected speech recognition. So, we are starting to give you a hand. We are doing something already and these are some of the directions of the re-

search lab. What we will be looking for is things to spin
out as they come out of the program. The problem is that
we don't have any spin offs. I've read that report too.
I looked at it on the plane coming down and what it says
is that you can't evaluate it until November, 1976. But
we think we have all these good things coming out of the
system. We, as R&D managers, have not gotten one thing
out of the system so far. Not one thing that we can real-
ly brag about. We would like to talk about it because it
would be support for you. So maybe one of the things that
we address ourselves to is what can we use from the system
as it stands right now? What kind of semantic modules or
syntactical modules phonetical rules that are going to
fall out of the system that we can use in our particular
scenarios.

Carlstrom: I agree with you on that. By the way, you
don't have to convince me about most of the things you've
said. I agree but I am glad that it is on the record any-
way. The problem is that I see all of that, but I am very
intimately involved now and emotionally involved in all
this and I see all those paths. The problem is how do
you find a champion? He is not even an R&D manager. He
may be a budget manager. We may have to go build the case
at some higher level. You have got to convince him that
this is an important area to do research. His model, is
that he sees a few successes and they can do that. He
doesn't have a finely structured picture of all these pro-
blems.

Beek: There is an option. You can either bury it, or
highlight it.

Carlstrom: What if we went the other way, and said, "We are going to discontinue research, it is solved." It is a solved problem. Maybe we should.

Goldstein: It keeps coming up that if we demonstrate the progress to R&D managers, they are going to say, "No more R&D money." There it is; "work it." However, if they would come down and see these systems, and if the operational people out in the field saw them, all would understand that it means just a little bit here and a little bit more needs to be done. There has to be a continuous recognition in digits. The "word" is okay for this portion, but they would see that the additional R&D funding would still be required. There would be that reinforcement that something does work. I feel bad that the argument is continually thrown out that we shouldn't show the R&D managers to soon.

Carlstrom: No, no! I am not saying don't show them. You have to show them. Just be very, very careful how you go about it.

Walker: There is another problem. It isn't a question of unwillingness to show them. We are trying to build systems that have a lot of complicated parts. I think all of us who are involved in system building would say that the overall success of the effort is going to depend on how these things relate to each other. We have had little leisure, and certainly no money, to look at the pieces by themselves in order to try and pull them out. We can say we are very comfortable about them. We can argue, I think, in sort of private conversations, that these things are real good and you ought to look at them. But we certainly

135

haven't had a chance to do the kinds of evaluations that would provide you with some creditability. Provide it in a way so that you could use the evidence to support your own R&D actions. You notice I am not saying that we are trying to be defensive, or secretive, or anything of that sort. We haven't really had a chance to evaluate the total system, or even to evaluate some of these pieces.

Beek: I understand your problem. But as these things become abailable, maybe you can get us out there and we can take a look at them. We might find some of that work desirable. We might do some of the funding on that. If you could identify pieces that would be interesting to us.

Carlstrom: I don't know where to go with this. All of a sudden we went through a lull, and now I am in a storm here.

Lakerson: I would like to suggest that we consider going through a phase of research kind of thing. One would say the next level, now that you have looked at the whole problem, would be a kind of a spin off. For instance, like the training situation. The next phase, of the development, would be more the connected speech. The next phase, after that, would be the mobile speaker situation. In attacking that, one would show the managers the forward stepping; walk before running. This would be a more orderly way to show the research that needs to be done.

Carlstrom: Of course, that was the risk at the beginning. The step by step approach versus the approach of taking on a total system. There are advantages and disadvantages both ways. The total system approach has caused most of the problems that Bruno points to. But, on the other hand,

it forces the people to be able to make the other kinds of
internal trade offs. It is interesting, I think, that part
of the final analysis will be not so much who wins the horse
race, but look back at the comparative advantages and dis-
advantages of the accomplished research. Compare the way
BBN made the trade offs versus the way SBC made the trade
offs. They picked slightly different internal structures
and knowledge representations. Neither one of them knew
for sure which one was the best way. They played hunches
and they placed their money on their horses. That is part
of the deliverable. It is to get back inside and sort
those issues out.

Woods: I would like to respond a little to the specific
question as to whether or not we might have spin-offs. A
couple of the things that I heard today are those spin-
offs. One very reasonable, and relatively short project
to do, in conjuction with the speech understanding program,
would be to put in a grammar continuous spoken numbers
group. That is the kind of thing one could do in the speech
understanding program. You could set it up so that you
could get exactly the kinds of performance numbers that you
would like to get. You could get a handle on the relia-
bilities, etc. The other thing that I think is quite a
reasonable spin off, is the issue you raised in your sur-
vey. That is to get some of the speaker variability things
into the program. Get some of the speaker ambiguity
things. You are going to have to do more of an analysis
at the phonetic level, and not to gestalt the whole pat-
tern of a word. Instead, identify the sub portions of the
word to correspond to different sounds. There is going to
be a lot of work on that. It is suggested that you try to
do a bit compression transmission, a vocoder type of appli-

cation, to get down to the 100 bits per second range,
where you are setting up the characters essentially. It
is the kind of an experiment that I'd like to do at any
of the current speech understanding sites. Pick a vocabu-
lary of first approximation where one can do a phonetic
analysis using the best pigmentation path. One should do
some feasibility analysis to figure out what other func-
tions or logic paths produce a neutral version of that
sound to synthesize at the other end. Is it a little bit
neutral as to which of the confusable sounds it sounds like?

Connolly: There are a variety of experiments of that sort,
that fit very nicely as small projects on top of the speech
understanding systems. They could be fairly large effort,
if you didn't have to use the facilities and body of skil-
led people, you have already put together in the current
speech understanding program. But as an add-on, the ex-
periments could be done on top of the things underway.

Conway: I would like to make one remark about the "many
speakers" question. I think it is almost a prime error.
I would like to assure you all, that as you fly home, we
don't let anybody off the street come in and control traf-
fic. It seems instead that some reasonable adaptation of
the man and his tool is essential and always will be.
That is all!

Little: I see the gentleman who was concerned about the
silent people has left now but we do have 4 people here
from the Bureau of Standards. We are here because we are
interested in speech recognition as a means of data ac-
quisition. We are also interested in voice verification
in the security field. The Department of Transportation

has hired us as sort of a technical arm to recommend appli-
cations in the transportation field, related to speech re-
cognition.

Meissner: I have been listening to some of the conversa-
tions here and maybe one of the things we should set our-
selves to doing at the Bureau of Standards is establish-
ing, and maybe promulgating, a standard American dialect.
I'll mention a couple of categories of our activities.
One is speech identification. The other is speech under-
standing. The identification people try to throw away all
the intelligence and just find that which is specific to
the speaker. The understanding people would like to eli-
minate speaker dependent characteristics and come up with
just pure intelligence. I can see an important applica-
tion to both and that is what Dr. Mundie was leading to.
We would like to have control, with the use of computers,
over access. Therefore, we would like to continuously
verify who it is that is addressing the computer. There
are other methods of identification. For instance, fin-
gerprints, which are a one-time thing. Having done that
it's not valuable as a further form of entry. But with
speech, you can do a better job of verifying the speaker
with some special selected training phrase or password.
But you can get some degree of recognition continuously
by addressing the speech understanding device, and we feel
that may be an important application.

Carlstrom: One comment. Joe Dixon from NRL mentioned
that they were looking at band with compression. The ARPA
programs overlap, except they stop at the LPC coefficient.
It is really a question of moving that petition further
along in the speech understanding system. Extending the

139

commonality. The contracting guy came in last night, who
will go unnamed, who is proposing this capability. He has
proposed to actually send speech phonetically at a 100 bits
a second over a teletype circuit. He very quickly passed
over the little box that was going to pull out the phonemes.
There must have been 50 boxes on this blocked diagram he
showed me but that one was very obscure. I had to really
look for it. I asked him how he was going to do the things
that went inside the box. He said, "Well, we haven't fig-
ured that out yet, and I am sure that is fairly easy to do."
It turned out that the reason he came to see me was because
Bob Con was out of town. It turned out more appropriate
for him to be talking to us because he is really talking
about building a phonetic detector for band width compres-
sion.

I still feel there were some hands that were
passed by in the last couple of iterations. If anybody has
been trying to get some comments on the floor maybe this
is another good point to do that.

Moore: I suppose most of this group is familiar with the
work IBM is doing in this area. I learned in a seminar
down at GW a couple of weeks ago that one of the applica-
tions they visualize is that instead of dictating a letter
to your secretary, you speak to a machine. It will come
back in soft display, with a first draft showing represen-
tation. You can point to words indicating no changes are
meant here. This is a tremendous application that every-
body realized. We have thought a lot about that. Your
people, I guess, are plugged in the closest to this pro-
ject.

Carlstrom: No, I don't think IBM is doing speech under-

standing; I think they're paralleling the lower processing
levels. In fact, they don't claim they are. They might
claim that their techniques are competitive with ours, or
should be tried in our system. They'll make that argument.
They wouldn't claim they had a speech understanding pro-
ject. Now the thing about IBM, which is sort of interest-
ing to me, is that they claim that htey do not use a limi-
ted knowledge domain to do their task but that's a little
bit unfair. They really do. And they definitely are con-
straining themselves to a dictionary on laser patterns,
etc. But then, having been a little bit hard on them, I
sort of feel that, gee, the electric typewriter, or the
speech typewriter can work because, as you pointed out, it
doesn't have to be perfect. If I could talk to the thing
and then go back with a light pen or something and scratch
the word and say it again in an isolated node and it would
fix it. That would probably be better than the through-put
I get right now.

Walker: My understanding is that IBM was not using speech
understanding. Unless they've changed since the last brief-
ing. They are very explicit about their being a speech re-
cognition system.

Reddy: They are very particular about that basically.
Eventually they want a voice typewriter and with unlimited
vocabulary. In fact, they want all of the English language.
I think you may be able to do it at certain accuracy levels.
Almost any system could do it, maybe about 50% accurate.
But, the question is, can you stand it? That is if you
take any one of the word hypothesis, used in the "under-
standing" systems. If you gave it a million words lexicon,
you would probably receive a thousand words as possible

candidates. Now you could build other verifiers, etc.,
which would probably reduce the system maybe ten or twenty
words. The question is, "Do we have enough knowledge in
the signal?" to reduce the ambiguity from 20 to 1. The
feeling here is there isn't. You just have to use higher
level knowledge. In other words, you're not going to make
it. Their hope is that they can actually do a large part
of this, not quite not using constraint, but using very
general statistical constraints about the language of Eng-
lish. That might help a little bit, but we don't know how
much.

Carlstrom: Well, my point was, I understand they claim
they're not doing speech understanding. I know that they
say that, but then they should use a random number genera-
tor to pick their vocabulary. Instead, they talk of using
laser patterns.

Reddy: They use vocabulary constraints and syntax con-
straints.

Woods: I think what they are getting at is that they don't
understand the sentence.

Carlstrom: I know, I know. Well, let me see. It's twen-
ty after four, and I don't have any great compulsion to
hold people here against their will, but, maybe we should
try to get into some summary phase. Try to wind down. I'm
open to suggestions but maybe the best idea is to have peo-
ple comment on what they think has been accomplished. What
is the message? What should we go away with? What should
the next steps be? Should we try to have another gathering
like this? Say six months from now, with just the Govern-

ment people present. Should we have another meeting like
this with open public and private participation? I'm open
to ideas. Maybe we can kick some of the things around a-
gain. Maybe we should go around the table and ask people
for their comments. Any thoughts, any first order thoughts
from anybody?

Neuburg: Well, I'm very ignorant about these management
things. Is there any other management, that would cause
a meeting like this to take place? Does anybody have cog-
nizance, or whatever it's called?

Carlstrom: Well, it would be nice if there were a cham-
pion other than ARPA. Not that ARPA shouldn't champion,
or hold another meeting. But, the thing that would help
the cause, I think now, would be for somebody in DDR&E, or
some other organization, to make a case for speech research.
A government-wide NBS can certainly do that, or the NSF,
or somebody like that. Or FAA.

NSA: If it is strictly government, and had dedicated NSA
interest, then we could be considered. If it was for an
open public thing, you'd really want a good turnout, and
someone like NBS would be more appropriate.

Clark (NBS): Well, I'd like to volunteer NBS.

Boehm (NSA): You would get very good turnout there for
side issues.

Carlstrom: Well, quite a few of thes people today came in
from quite some distance. San Diego, I know, is here and
where else? I mean apart from the ARPA people. There

are a lot of Navy people who seem to have come in from quite
a ways. I guess what I'm concerned about is that it would
be nice if somebody else could, besides ARPA, make a strong
case. ARPA obviously has an investment, a vested interest.
We're trying to handle that interest in a responsible way.
The thing that would help my case though, would be if other
groups, and DOD collectively, or through a common champion
said, "Hey, this is an important problem." I think there's
a built-in assumption that, "Well, don't worry about that;
ARPA is funding it." The darn momentum of bureaucratic
problems is a concern. The word doesn't get out until we're
one year past the funding and the budgets are firm. We
plan so far into the future that there's quite a span be-
fore other people can pick up the ball. I guess I'm sort
of encouraging people that feel this is important to try
to champion the cause. It's also conceivable that my own
management will reassess its priorities if it views this
problem as being desired by somebody. And, let's face it,
they are pragmatic, too, in establishing priorities. If
they look around the Pentagon, and they look over in DOD,
and they don't see anybody that gives a damn, they find it
pretty hard to go to the Hill and justify two or three mil-
lion dollars. It's very, very hard if they can't back it
up saying, "The following five agencies are screaming for
this kind of technology."

Bernstein: Is six months too long a time? Because if we
wait six months we'll be beyond the budgeting cycle. And,
too, what will be the content of the next meeting if there
is to be one? Shouldn't we address the technology issues?
Or should we address the issues on "what we really need?"
Should we forget about where we really are? Not forget
about it, but from the point of view of the vector of the

144

meeting, should the primary focus be what are the needs? How should we get everybody at least focused in one general direction, rather than going off in ten different directions? Re the third question: What sort of participation do we want from industry?

Carlstrom: I seems to me that there needs to be at least two kinds of meetings. One, a meeting of the public domain which may go the whole nine yards. Here one would make an announcement in the Commerce Buisness Daily, or something like that. Maybe just hold a two day session with all the players there. But then, there also needs to be a meeting of maybe just Government people to talk about requirements and budget justifications, etc. Because, if somebody in the Government can't stand up and really make a strong case for the need for these things, it isn't going to happen. Everybody knows that industry wants the business, right?

Beasley: Right now, you are 44 hours late on that budget cycle. Today is my "drop-dead day" to be here. I worked all night because I had to complete my full '77/'78 program. It had to be identified and marked out. The '77 budget went to bed last night. But, as far as our government interest goes, there are two other government users represented here that have as much interest as I have. I have four funded projects right now, two of which are bio cybernetics oriented, but it's only an interface package. It is not total system application. And so, I would say that we are definitely interested in some advance in the state of the art. We want the latest things, naturally. I would say that it would be very appropriate for us to help you put on, or sponsor in some way, the next meet-

ing. But, one of my very promising systems, that would
use this, is a Navy system that we're funding very, very
heavily. This is in the next year and the year following.
It could dangle over to this field. But that's something
I have to go back to the project officers and redefine
some objectives. So, it's not something that I can discuss
from the contractor interest at this time.

Carlstrom: As far as the lead-time problem, ARPA didn't
perceive this situation, as a crisis situation, until a
few months ago. It was internal ARPA management. In the
past, we hoped to approach the final demonstrations in a
reasonable way. As we near those demonstrations, we will
talk about what will happen in a follow-on. We'll probab-
ly bank some money to do that kind of thing. If it doesn't
look like worth doing, that money could be used to do other
things. Now the issue is that management has decided that
these things aren't worth doing. Even if the "show and
tell" walked on water. You know, they've sort of already
said, "No sense waiting to see those results 'cause we just
don't think this was an important area." So, it's impor-
tant to try to build a very strong case. Now as the case
is made, through various DOD channels and etc., I don't
think these views can't be reversed. Because I think all
people involved are intellectually honest about what's
going on; it's just a question of finding priorities.
What are the priorities? Nobody has enough money or enough
resources. If the priority for this sort of work is high
enough, dollars can be programmed in proportion to those
priorities.

Dattilo: Are you indicating, or are some of the other
people indicating, that you would want further input as to

system requirements, from the other DOD electro research people?

Carlstrom: Absolutely!

Dattilo: Because we have, as far as the Army's concerned, a laundry list. We are very emphatic on this. We would dedicate our own word recognition system for the field. I could go as far as to say that success of the TOSS system might depend on the success of the word-recognition system. Because that's the way the Commanders feel in the Army. They want to speak into the system and get the data output. They all want to go through digital devices. They all want to go through machines to do it. So I think the Army urges you on. We could give you input from our plans. You could help us, really. We could find out where we stand technically. Because, from that base we generate requirements. The requirements will follow technology. Not that we generate requirements that are "pie in the sky." We try to keep requirements in line with what technology can give us within a certain time frame. So it would be helpful to us to have this kind of dialogue going back and forth.

Beek: Yes, he's really hit the nub of the problem with all of us. And that's really the nub of the problem! The requirements don't determine the technology. It's usually the technology that determines the requirements. And you know, Dave, in the military that's the way the real world is. So, we don't know what you can do; we can't generate these requirements. The only way that we can get the requirements is if we more or less guarantee that we can get positive results. Not abstract, but positive results.

147

It's very difficult. R&D money, 6.2 money, in the past,
has been very dear. 6.3 money, when it comes to building
equipment, seems to be kind of loose. People seem to have
lots of money for that sort of stuff. But the R&D money,
for basic research, (and 6.2 money) is scrutinized under
a microscope. So what we have to do is to be pretty well
satisfied, in our own mind, that we can make some inroads
in these programs. So, I know your problem. You'd really
like us to come up and say, "Here's ROC such and such. We
don't; there's no military requirement for speech under-
standing and there won't be. I can almost guarantee you
there won't be until we really know fairly competently
that we're able to do it.

Carlstrom: Well, I understand what Doug's problem is--you
can fingerpoint between the user and the R&D guy. Each guy
blames the other. But it's really a circuitous thing.
There has to be a lot of interaction. Each guy can help
the other guy out by stairstepping. But I agree with you.
I understand what you're saying. What that really says is
there should be more meetings of this kind. It also points
out another thing that we've come across in the past. It
is that it's not enough for ARPA to get with service labs,
because they're having the same problem we are. You have
to occasionally drag in the end user. You have to bring
in the fellow who really writes the ROCs. RADC doesn't
write ROCs. My view in the Air Force, though, is that the
R&D people are too paranoid about that. They're hurting
the ROC process by being too defensive. You might get
zinged occasionally, but there should be more interaction.
But, I'll basically agree with you. It seems to me that
more meetings like this might help solve that problem.
It's not clear to me that it would, because if the R&D

guys just get together, and talk to one another, without
ever bringing in any end user it won't work. Now, you're
being involved with photo people and you're obviously out
there trying to get in bed with the end user to help justi-
fy your case. I guess that's all I'm saying. It seems
along those lines, the issue is to have meetings like this
more often and the indication seems to be we should have
been doing this for the last two or three years. At least
every six months having a meeting of this sort.

Walker: Well, as a defense of view, and of ARPA, I think
it's fair to say, it would make visable more promises.
You know, six months, a year ago. I think some of the peo-
ple out here would have been more willing to tolerate our
expectations. Two years from now, we will be talking about
some things when we have systems operating. So it's sort
of a different thing. I think it's very reasonable to have
this meeting now. It could have been earlier, but I'm not
sure how many of the other kinds of capabilities that I'm
hearing about for the first time. The Navy exercise, and
Wherry's work. Now it seems to be sort of coming together
rather close to practical application.

Carlstrom: Well, it seems that there are two issues. One
is the critical mass, quantum jump, loss of R&D versus tak-
ing small steps where you maintain very good scientific
control. The argument is, if you take small steps in cer-
tain classes of problems, by taking small steps you'll ne-
ver solve the problem. Because you never get control over
all the variables necessary to solve the problem. But,
even so, I would maintain that having, maybe not every six
months, but say, annually, interactions like this would
have put the program in a stronger position. We could be

doing essentially the things we're doing now. But some of
our discourse domain might have been more closely tied to
some of the analogous problems that people in the services
are concerned about. If for nothing else, but just politi-
cal or psychological, that fact would have been very good
and useful. We feel fortunate, I think, that the SDCS sys-
tem did get plugged in closely with NELC. But, I think
there were more opportunities for that kind of thing. We
missed out on some of those opportunities. An earlier ses-
sion like this might have captured those. But, on the oth-
er hand, there is the legitimate issue. I guess it's sort
of like what you were saying, Bruno. If you could get too
involved with the user it's hard to break away and do any-
thing. It's hard to get out of his day to day concerns.
Well, let me rephrase that. I'm not trying to get people
to just buy the ARPA work. What I'm trying to do is get
people to justify speech research. And the devolvement of
the technology across the board. Regardless of the succes-
ses and sins that have occurred anywhere. OK? ARPA has
incurred some sort of responsibility for the devolvement
of the entire technology base, whether or not it's repre-
sented in the ARPA program. I think, I'm really asking
for your help in trying to make a case for speech research.
I guess what I'm thinking is, if we can get some kind of a
formal official organization, something like the existing
speech consortium, devoted to speech recognition. Not
speech understanding, nor word spotting, but just to speech
recognition research. See, in the speech consortium,
there is a nice situation. They have a champion. They
have someone in DTACCS that call the meetings, establishes
deadlines, and manages the thing. It would be nice if we
could get some similar group together because that tends
to produce a lot of quality. Then have a meeting and you

can have a spokesman. You can go and talk to people, per-
haps go talk to Dr. Currie. Go talk to other people and
make a case. I guess what I see is really required is set-
ting up something like that, getting something like that
established. Is that feasible? We can go along and have
these informal meetings but all we'll do is start more and
more agreeing with one another, or at least, agreeing to
disagree. But what we really ought to try to do is get
some kind of a consortium.

Neuburg: That was what I was trying to suggest earlier.
There has to be a mechanism of some sort.

Medress: It seems to me that the answer to that question
probably becomes a new perception of the technology that's
being devolved. I think Bill tried to make the point that
there are a lot of experiments that could probably be done
quite easily now because ARPA has made the investment, to
build the system and has put the staff together. One of
the things that people are concerned about is that it might
all disappear. The potential payoff that should come may
never be realized. That's one of the reasons for us to
get together today. And, if that's perceived as such,
that motivation disappears.

Carlstrom: I don't know what it takes to get something
like a consortium established. I know these things exist
but I don't know how they get started. I mean it's clear
to see how they get started if some high level guy perceives
the problem and does a top-down. They get started very
cleanly and quickly. Maybe the thing to do is get involved
in the DDR&E Study Committee...They asked me to be on that
and I said, "I don't know anything about speech. I don't

151

know why you want me on it." I asked them to go back to Dr. Heilmeier to ask the office for support. This was done to deliberately get it through formal channels. Maybe that's the beginning of something like this. It's a much broader thing, it's just like one chapter or something being devoted to speech.

Beek: We, too, have requested the coordination of practically everyone, except the Navy. The Navy is the one we're having the most difficulty with. We have coordination by NSA, also the Army. But we have in our own set, a method of coordination. We try and, because of the speech, we're also coordinating with ARPA. There is already an informal structure for doing this.

Carlstrom: What drove the NATO thing? I mean there is something that seemed to happen spontaneously.

Beek: Well, Ned said he had to drop out and talked me into it, because we've worked very closely together. That's how we work and so we know pretty well what one another's doing in this area. I think having a government meeting, a closed meeting, would probably help everyone. Especially having a closer tie-in with what some the Navy is doing.

Carlstrom: Well, that seems to be it. I know everybody has to go. I'm worried just trying to capture the important points at the end. It seems to me that an action item is to have a subsequent meeting. Hopefully, one within a couple of months. It should include all the government people that can possibly attend. It also seems like a nice idea to have a public meeting at NBS. I don't know what the issues are on doing something like that, but it seems

to me that there ought to be two subsequent meetings. One in the public domain to talk about this and one meeting of government people to talk about how to coordinate and build internal mechanisms. Would anybody volunteer to try to put together either one or both of those meetings? Or do the people agree with me that they're both good ideas?

Suttle: As the government I guess we sort of have two problems. There is the DOD problem and there is the government problem. But in addition there is the public problem. Are you suggesting that there's no, as far as I'm concerned or he's concerned, there's no difference between the government problem and DOD's problem. Are you suggesting the organizations be only DOD people?

Carlstrom: Well, I don't see any unique internal DOD problem. You just can't have a government meeting and get everybody there at once. Maybe the DOD guys have a separate session or something if they want it.

Suttle: We have a broad mind over in our shop where we consider all government people over there as DOD employees.

Carlstrom: Alright, I guess I would really vote for a government-wide meeting rather than just a DOD meeting.

?: We definitely want the other government elements in because they're more concerned really in this particular area right now than I think DOD is. We want the fringe benefits of their money.

Carlstrom: Yes, I would say it's in DOD's interest to include non-DOD government people. Well, what I've had in

mind was that there is a possibility of getting leverage
or non-DOD money. It's becoming increasingly important be-
cause it's harder and harder for DOD to be the shoulder for
basic research in this country. Somebody else just has to
pick it up-I mean NSF or somebody has to. You know, one of
the reasons DOD's getting hard-nosed about it is because
Congress gives us a hard time about supporting basic re-
search. If it's good science for science's sake, they say
that's what the National Science Foundation is for. That's
the only reason I would not like to exclude the non-DOD
people. But I also like you comment, that it is a little
more comfortable and easier to talk in terms of DOD pro-
blems, if it's just a DOD meeting. I guess what I thought
could happen would be to have a day where there's a govern-
ment meeting and then, in the afternoon, or something, or
as an auxiliary to that meeting, have the DOD people go off
and have their own meeting. I don't know.

Suttle: Well, the only reason I make this comment is be-
cause I've had some experience in research management, the
same kind of experience you've had. Each year we have to
justify basic research programs. Our research office has
only basic research programs.

Carlstrom: Right.

Suttle: It is important that all of the Army offices, the
Army Research office, was the only office to have a budget
increase to the tune of 50% for next year. This is be-
cause of our briefings to your present cause. Last year
we didn't get anything. Our recommendations were limited.
So, we've had some experience in doing this "show and tell"
for our research programs and have some knowledge of how

154

these things are done.

Carlstrom: Well, I think though, a lot of the problem is
outside of ARPA. I think there's Congressional concern
and critical review. Rightly so, one must make sure that
the various programs are all in. So, it's not really just
an internal DOD problem. It is broader. One problem that
our management is really worrying about is to make a solid
case behind every research dollar that they have to go to
the Hill. They must make creditable arguments as to why
the programs are important to DOD. Well, again, who would
be willing to sponsor the meetings? Either kind of a meet-
ing, a DOD meeting, a government meeting, or a public sec-
tor meeting?

Clark: Let us sponsor one and government sponsor one.
We've done this before for DCA, for example, giving them
our facilities for them to hold a system telecommunication
program. This type meeting would be the most appropriate
for us to hold.

Carlstrom: Okay. I think you can tell I'm trying to dodge
the responsibility of pulling the next meeting together.

Suttle: I believe I can talk the DNA into jointly sponsor-
ing with you and hosting at our facility a government meet-
ing on the subject. The reason I don't think I will get
the same reception to a DOD-only, is that I don't think I
can go that route. But we are on joint committees and we
are in big with ERDA and NRC all the time. We are the staff
element there for the Assistant Secretary for Atomic Ener-
gy. Being that we kind of work halfway on defense. We do
as much work for NRC and ERDA. Being the liaison between

them and the military department we find it rather unpolitic to exclude ERDA and NRC on a subject as vital as this. So, if we were hosting it, I would have a lot of daggers thrown at me if I couldn't invite NRC, NBS, and ERDA. Under those circumstances I wouldn't volunteer to host it. We have money for such a meeting as this. It would be mainly to find out who is doing what, and where, and how much. What's going on and that sort of thing. Just general, for official use only level. Not a classified level. Then, whether or not following that meeting, it would be desirable to get half a dozen people together who is really in the 6.1, 6.2 arena, from the three services and the joint services. I think that might be very desirable and I'd endorse that. We could talk about how we're going to put in our two cents worth. But I want to remind you all that I'm a peanut merchant--you give me a little bag of peanuts and I go around and feed the monkeys. I haven't got any money...that's all upstairs in my shop.

Carlstrom: Well, I like the idea of NBS holding a public sector meeting. It somehow just seems right, from a lot of points of view and that we somehow work out the government DOD-meeting somewhere else. Now would you [Suttle] be willing to take that on your watch to try to take charge of something like that and run something like that? I think that would be important. It is a worthwhile thing and we can talk some more about it later.

Beasley: Can I put in a comment about the public meeting (that might have industry in.) There was a comment made earlier, a suggestion which was good, that if you're going to talk about different requirements, have somebody specify them. So it sort of like it was done in the Newell Re-

156

port for the ARPA speech understanding system. Specify the dimensions along which you'd have certain conditions, so that you can look at various requirements and decide which of these might be the appropriate ones to begin with in subsequent research.

Wayne: To define, precisely! This is one of the greatest things that came out of the ARPA people working together in the rules workshops and things like that. Let's have all our rules specified in the same context and then we can look at them and compare them to each other. In the same sense, if you could get some systematic and fixed way of specifying applications, so you could see some of the dimensions of them and compare to the select ones that are appropriate. That would be very good. In fact, if you could do it before you had a meeting, then you could come in with that kind of information to interchange.

Carlstrom: I guess I have to agree with all that. But, the main thing I'm interested in is getting somebody to agree to make that happen. It becomes very hard. It's just like a lot of the wrinkles that happened in this workshop. I know a lot of things should have been better coordinated ahead of time but it's just really difficult to take the resources in terms of one's own time. I feel lucky it went as smoothly as it did today. Again, if NBS would be willing to try to take upon themselves the task of putting a public sector meeting together, I would really appreciate that and would try to do what I could to help. This issue of a DOD meeting still seems very important. I don't know, I guess I'd ask Jim Suttle if he would take that on his watch for the time being at least, try to figure out how we could do that. I'm just afraid if I take

all these things on my watch they won't get done. The issue of having a government meeting is still appropriate.
Maybe we can hold off on that. Okay. I guess maybe that's a reasonable enough compromise at this point. I really think that a government-wide meeting is bound to happen somewhere between these two extremes. I'll just leave that and maybe we can stop at that point. I just wanted to make sure I could get somebody to agree and I think that we should agree that those meetings should take place in the 3 to 6 months time span, the sooner the better. But I know it takes time to, especially the public sector meeting, to really put something like that together. So, it's probably going to be closer to 6 months, and maybe that's fine. Maybe the meetings that have to occur quickly are the government people meetings. Then they can start trying to get these budget issues in order. Get some kind of coordination and get some reason out of a potential chaos that sits there in the budget structure. So, I guess the first priority should be for Jim and I to talk and try to figure out ways to/of sponsoring a DOD meeting. Maybe it'll end up that either you or I do it. The Washington area would probably be the best. Consider those closed items. I know everybody wants to go. I know I do. Are there any other items? If anybody would like to send in their comments or critiques of this activity today, I would welcome that.
Both positive and negative feedback are very welcome. In fact, previously we included some feedback in the minutes.
If some of these kinds of comments came in and were received in time we might include them in the minutes as an appendix. That is, people's thoughts and reflections. Sometimes when people go home and sleep on what happened for a day or two, they have insights, etc. that are very valuable.
So, if people do send in, either to myself or Lee Bauman at

158

SAI, we will include them. If you don't want them published
just as afterthoughts, we can do that, too. At your discre-
tion I will include them as an appendix to the minutes which
will be mailed back out. If everybody is signed up on the
roster, you'll get copies of what happened today. Please
try to get slides for you viewcrafts to Lee Bauman to help
him with his notes. If you have to send those in later,
fine. Anybody else have any comment? I would just like
to thank you for all your efforts. I think it's really
very helpful to have all those who are doing basic research
together so that we know where we stand.

ATTACHMENT 1

ATTENDEES TO THE ARPA SPEECH UNDERSTANDING WORKSHOP
13 NOVEMBER 1975

NAME & ADDRESS   TELEPHONE NUMBER

Mr. Lee S. Baumann   (703) 527-7571
Science Applications, Inc.
1911 No. Ft. Myer Drive, Suite 1200
Arlington, Virginia 22209

Mr. Marvin C. Beasley   (202) 325-7395
Defense Nuclear Agency
Code ISNS
Washington, D.C. 20305

Dr. Bruno Beek   (315) 330-3454
Rome Air Development Center
Griffiss AFB, New York 13441

Mr. M. I. Bernstein   (213) 829-7511
Systems Development Corporation
2500 Colorado Avenue
Santa Monica, California 90406

Mr. John Boehm   (301) 688-8147
National Security Agency
R-54
Ft. George G. Meade, Maryland 20755

Dr. Robert Breaux   (305) 646-5130
Naval Training Equipment Center
Code N215
Orlando, Florida 32813

Major David Carlstrom, USAF   (703) 694-5037
Advanced Research Projects Agency
Information Processing Techniques
  Office (ARPA/IPTO)
1400 Wilson Boulevard
Arlington, Virignia 22209

Dr. Donald O. Christy   (714) 225-6515
Code NELC (3210)
U.S. Naval Electronic Lab Center
San Diego, California 92152

                                                 Attachment 1

1

NAME & ADDRESS	TELEPHONE NUMBER

Dr. Donald W. Connolly
ANA-230
U. S. Department of Transportation
FAA/NAFEC
Atlantic City, New Jersey   08405

(609) 641-8200

Mr. Franklin S. Cooper
Haskins Laboratories
270 Crown Street
New Haven, Connecticut   06510

(203) 436-1774

Mr. William P. Dattilo
AMCPM-TDS-SE-SDE
Fort Monmouth, New Jersey   07703

(201) 531-4159

Dr. John K. Dixon   *Ted Var*
Naval Research Laboratory
Code 5407
Washington, D. C.   20375

(202) 767-3851

Mr. James W. Forgie
Lincoln Laboratory
Massachusetts Institute of Technology
P. O. Box 73
Lexington, Massachusetts   02173

(617) 862-5500

Mr. James W. Glenn
Scope Electronic, Inc.
1860 Michael Farraday Drive
Reston, Virginia   22090

(703) 471-5600

Dr. Gordon D. Goldstein
Information Systems Program
Office of Naval Research
Ballston Center Tower #1
800 N. Quincy Street
Arlington, Virginia   22217

(202) 692-4302

Mr. Ira Goldstein
Naval Training Equipment Center
Code N-215
Human Factors Laboratory
Orlando, Florida   32813

(305) 646-5130

Attachment 1

2

NAME & ADDRESS	TELEPHONE NUMBER
Mr. Graham L. Gross LTS Corporation 71 W. 23rd Street New York, New York 10010	(212) 741-8340
Mr. Fred Healey Code 714 Goddard Space Flight Center Greenbelt, Maryland 20071	(301) 982-5683
Dr. David C. Hodge AMXHE U. S. Army Human Engineering Laboratory Aberdeen Proving Ground, Maryland 21005	(301) 278-3126/4389
Dr. June Shoup-Hummel SCRL 800A Miramonte Drive Santa Barbara, California 93109	(805) 965-3011
Mr. Rodney W. Johnson Naval Research Laboratory Code 5403D Washington, D. C. 20375	(202) 767-3012
Mr. Joseph J. Kalinowski Scope Electronics, Inc. 1860 Michael Farraday Drive Reston, Virginia 22090	(703) 471-5600
Dr. Wayne A. Lea MS VOP16 Sperry Univac DSD P. O. Box 3525 St. Paul, Minnesota 55165	(612) 456-2434
Major Leon Lake DCEC/DCA Code R740 1860 Wihle Avenue Reston, Virginia 22090	(703) 437-2474
Mr. John L. Little Room B212 Technical Building National Bureau of Standards Washington, D. C. 20234	(301) 921-3723

Attachment 1

NAME & ADDRESS	TELEPHONE NUMBER
Mr. Donald C. Lockerson Goddard Space Flight Center Code 714.3 Greenbelt, Maryland 20771	(301) 982-5378
Mr. Mark Medress Sperry Univac DSD M.S. UOP16 Univac Park P.O. Box 3525 St. Paul, Minnesota 55165	(612) 456-2430/2447
Mr. Paul Meissner A219 Building 225 National Bureau of Standards Washington, D. C. 20234	(301) 921-3427
Dr. Paul Mermelstein Haskins Laboratories 270 Crown Street New Haven, Connecticut 06510	(203) 865-6163
Mr. R. T. Moore Section 650.01 National Bureau of Standards Washington, D. C. 20234	(301) 921-3427
Dr. J. R. Mundie AMRL/BBN Wright-Patterson AFB, Ohio 45433	(513) 255-3673
Dr. Edward P. Neuburg R-5 National Security Agency Fort Meade, Maryland 20755	(301) 688-8147
Mr. D. R. Reddy Carnegie-Mellon University Pittsburgh, Pennsylvania 15213	(412) 621-2600
Mr. H. B. Ritea Systems Development Corporation 2500 Colorado Avenue Santa Monica, California 90406	(213) 829-7511

Attachment 1

4

NAME & ADDRESS	TELEPHONE NUMBER

Ms. Ann Robinson
Artificial Intelligence Center
Stanford Research Institute
333 Ravenswood
Menlo Park, California 94025
(415) 326-6200

Mr. Gerald C. Schultz
TST-48
U. S. Department of Transportation
Washington, D. C. 20590
(202) 426-4241

Dr. James R. Slagle
Naval Research Laboratory
Code 5407
Washington, D. C. 20375
(202 767-3850

Dr. Jimmie R. Suttle
Army Research Office
Electronics Division
P. O. Box 12211
Research Triangle Park, North Carolina 27709
(919) 549-0641

Dr. Donald E. Walker
Stanford Research Institute
333 Ravenswood
Menlo Park, California 94025
(415) 326-6200

Commander Robert J. Wherry, USA
Naval Air Development Center
Code 402
Warminister, Pennsylvania 18974
OS2-9000

Mr. William A. Woods
Bolt, Beranek & Newman, Inc.
50 Moulton Street
Cambridge, Massachusetts 02138
(617) 491-1850

Attachment 1

5

ATTACHMENT 2

AGENDA

SPEECH UNDERSTANDING WORKSHOP

0830 - Welcome and Introduction
          (Major Carlstrom, ARPA)

0845 - Review of ARPA Speech Understanding
          Program

1015 - Review of other Research Programs

1215 - Lunch

1315 - Discussions of Issues:

        ● Future Research Needs in Speech Recognition

        ● Inter-relationships of Word-Spotting,
          Isolated Word Recognition, and Speech
          Understanding

        ● Future Investment Strategies in Speech
          Understanding

1645 - Closing Remarks
          (Major Carlstrom, ARPA)

Attachment 2

AGENDA

SPEECH UNDERSTANDING WORKSHOP

0830   -   Welcome and Introduction
(Major Carlstrom, ARPA)

0845   -   Review of ARPA Speech Understanding
Program

1015   -   Review of other Research Programs

1215   -   Lunch

1315   -   Discussions of Issues:

- Future Research Needs in Speech Recognition

- Inter-relationships of Word-Spotting, isolated word recognition, and speech understanding

- Future Investment Strategies in Speech Understanding

1645   -   Closing Remarks
(Major Carlstrom, ARPA)

ATTACHMENT 3

ADVANTAGES OF SPEECH
AS A MAN-MACHINE
COMMUNICATIONS CHANNEL

1.   Most effortless encoding of all human output
      channels.

2.   Higher data rate than other output channels.

3.   Preferred channel for spontaneous output.

4.   Does not tie up hands, eyes, feet, or ears.

5.   Can be used while in motion.

6.   Can be used in parallel with other channels
      or effectors.

7.   Broadcast over short ranges.

8.   Inexpensive terminal equipment.

Attachment 3

SPEEDS OF VARIOUS
HUMAN OUTPUT CHANNELS

1.	Reading out loud	~4	words/sec
2.	Speaking spontaneously	~2.5	words/sec
3.	Typing (Record)	~2.5	words/sec
		(~5	strokes/sec)
4.	Typing (Skilled)	~1	words/sec
5.	Handwriting	~.4	words/sec
6.	Hand printing	~.4	words/sec
7.	Telephone dialing	~.3	words/sec
		(~1.5	digits/sec)
8.	Mark sense cards	~.1	words/sec
		(~.5	digits/sec)

Attachment 3

2
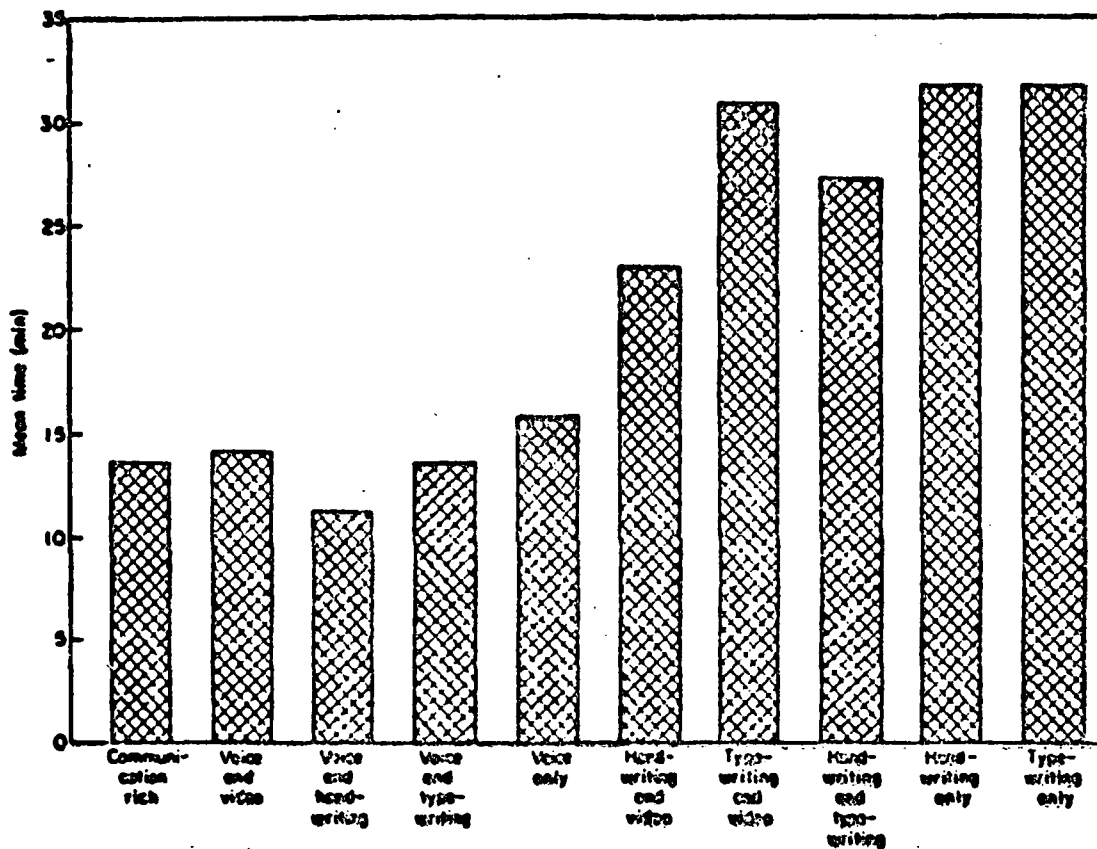
Fig. 2. Mean problem solution time for the 10 modes of communication.

3

TYPE B/65 SONAGRAM® KAY ELEMETRICS CO. PINE BROOK, N.J.

4
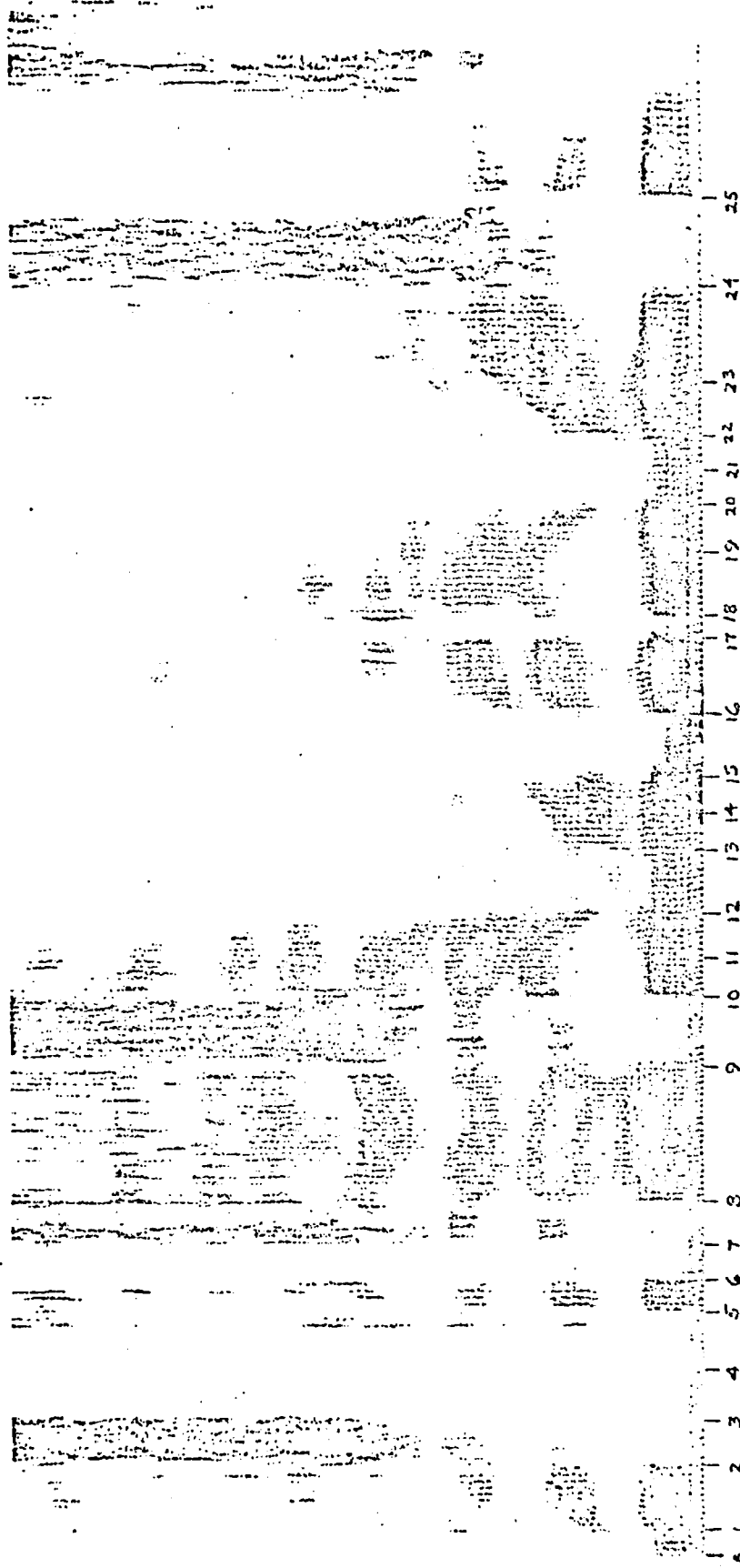
Attachment 3

0	(or L W)
1	Front V
2	(Or S Z
3	(And -voiced plosive)
4	(Or (and -voiced plosive) DH)
5	(And front v (Not IY))
6	(And -voiced plosive)
7	(Optional S)
8	(And Front V -high)

.

.

.

0   (L   W)

1   (IY IH EY EH AE AX)

2   (S   Z)

3   (P T K CH)

4   (P T K CH DH)

5   (IH EY EH AE AX)

6   (P T K CH)

7   (OPT S)

8   (EY EH AE AX)

•

•

•

Attachment 3

6

PHONETIC TRANSCRIPTION ACCURACY FROM SPECTROGRAMS

	KNS	DHK
SEGMENTS CORRECTLY TRANSCRIBED	24% ⎫ 74%	40% ⎫ 70%
CORRECT BUT PARTIALLY SPECIFIED	50% ⎭	30% ⎭
ERROR IN AT LEAST ONE FEATURE	15%	19%
SEGMENTS MISSED	11%	10%
SEGMENTS ADDED	—	1%
TOTAL NUMBER OF PHONETIC SEGMENTS	299	283
TOTAL NUMBER OF WORDS	80	67
% WORDS CORRECT WITH COMPUTER PROGRAM AS AID	96%	96%

Attachment 3

7

DWD-20

Attachment 3

Segment Lattice

Attachment 3

ten | been | did | give
we | give
are
magnetite
lead | been
not | did
and
are | done
did
done

with
give

s | sample | glass
are
moon
mode
lunar | sample
does
dash
percent
less | had

people
greater | dealing | metal | nickel
any
anyone
and
greater | dealing | metal | nickel

Attachment 3

11
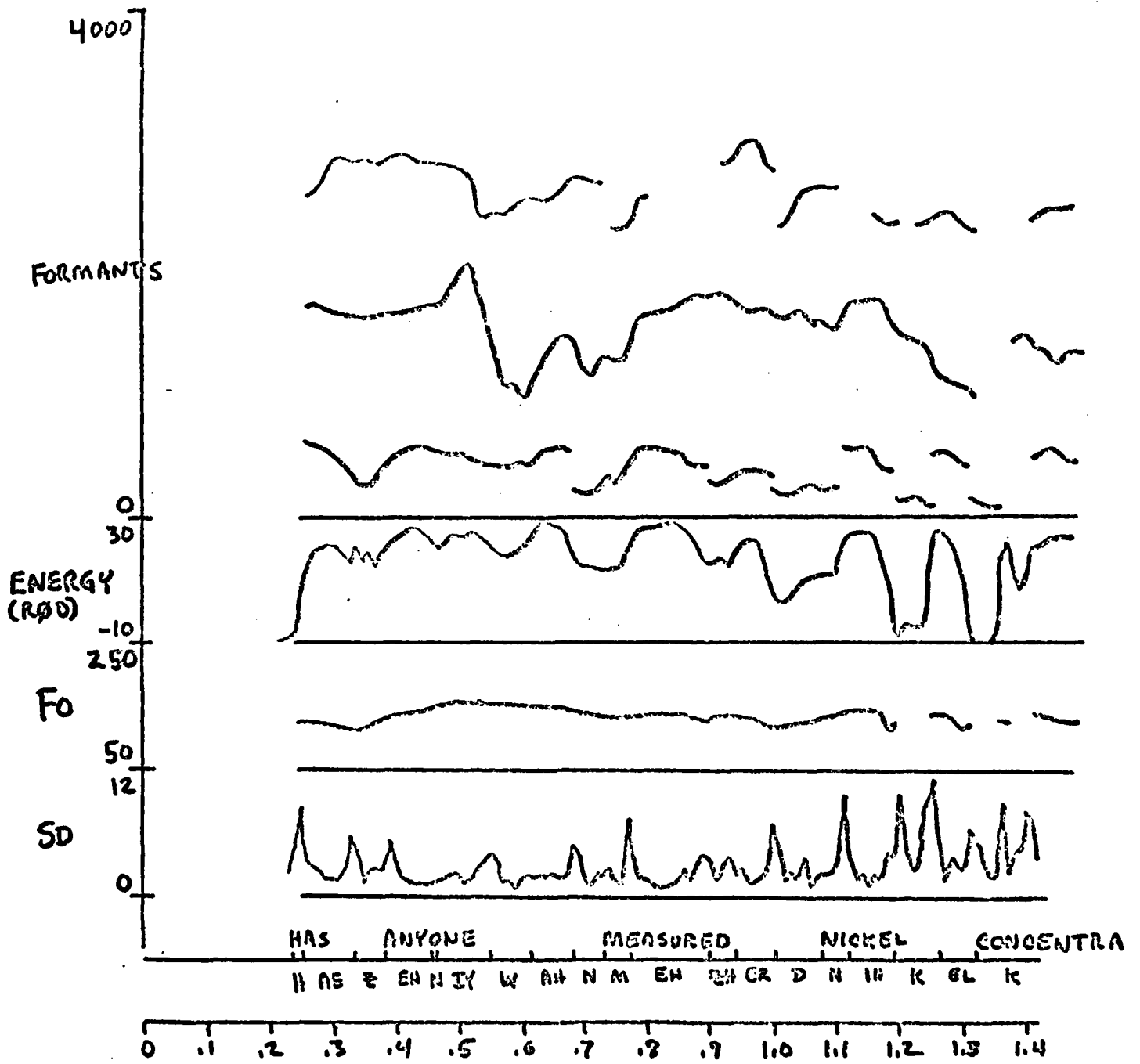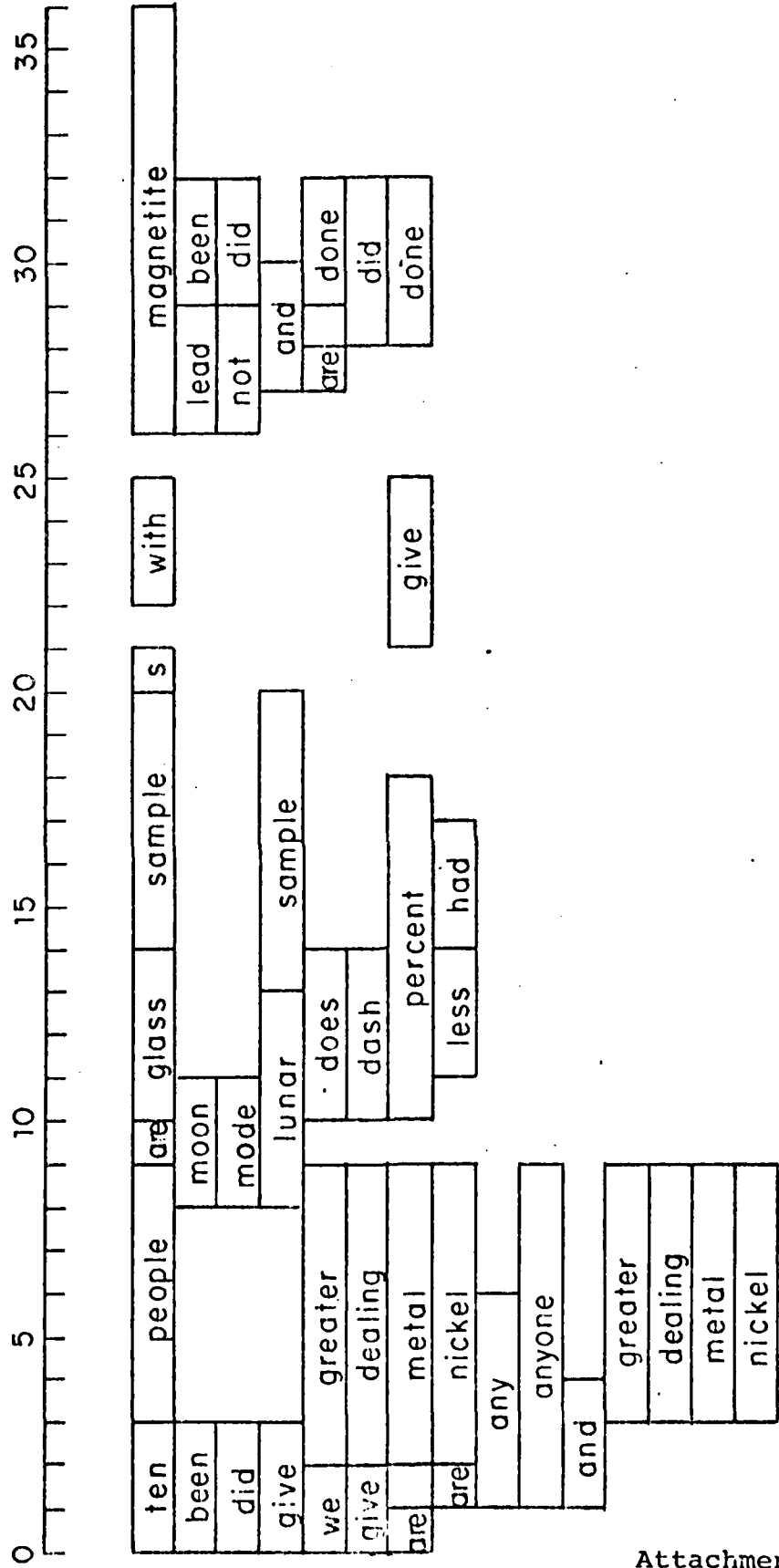
COMPONENTS OF A SPEECH-UNDERSTANDING SYSTEM
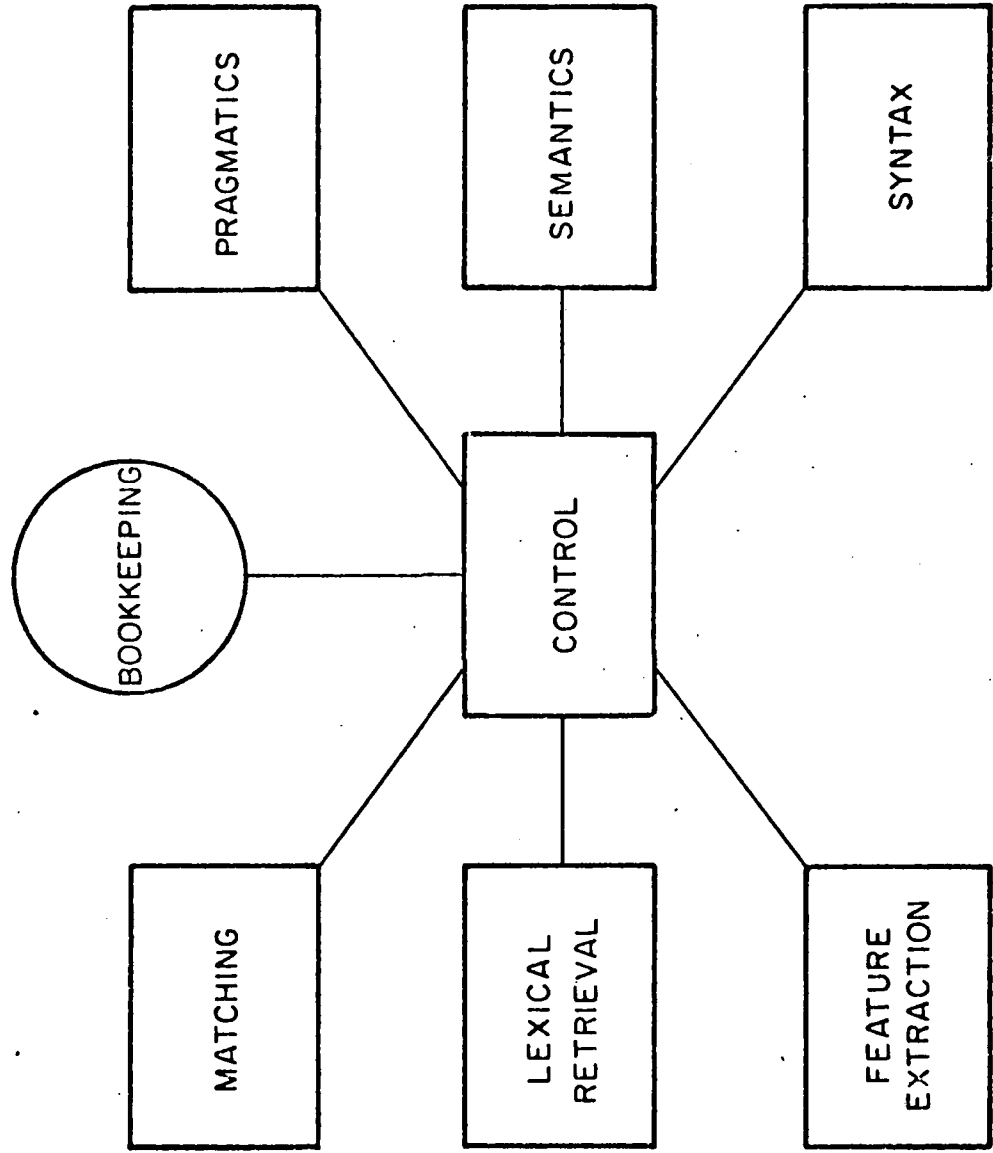
Category I

Closed Vocabulary
Finite State
Small Branching Ratio
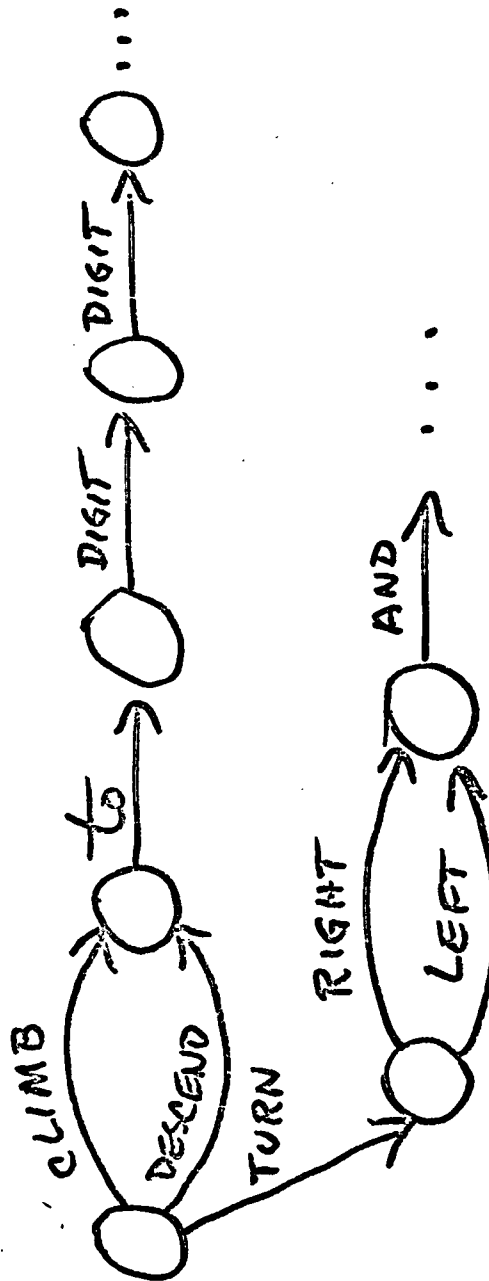
CATEGORY II

OPEN CLASS VOCABULARY

FINITE STATE GRAMMAR

VARIABLE BRANCHING RATIO
(LARGE IN PLACES)
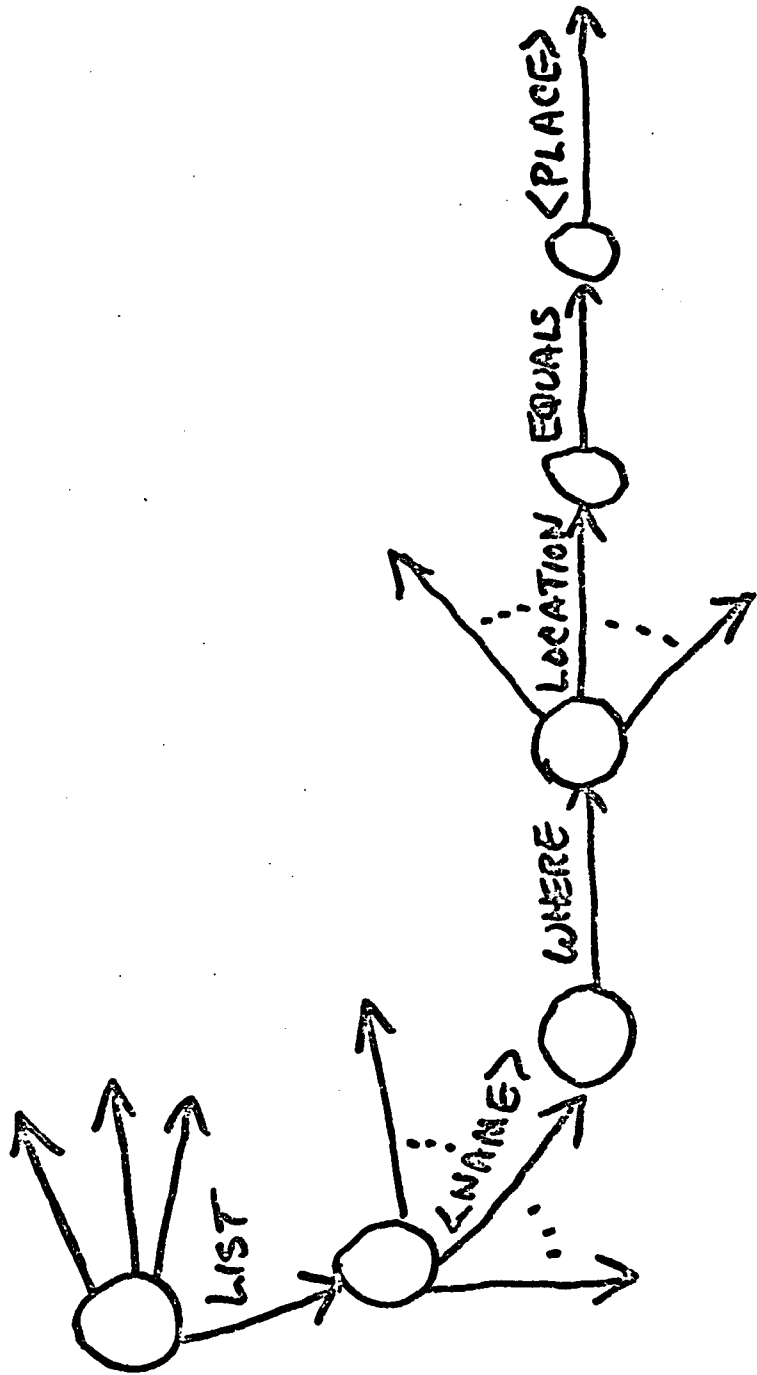
CATEGORY III

CONTEXT FREE GRAMMAR

PRODUCTIVE SET OF POSSIBLE THINGS TO SAY

$\langle$ UTTERANCE $\rangle \rightarrow \langle$ BASIC COMMAND $\rangle$ |

FOR $\langle$ VARIABLE $\rangle$ IN $\langle$ CLASS SPEC $\rangle$
(WHEN $\langle$ WHEN-SPEC $\rangle$)
DO $\langle$ BASIC COMMAND $\rangle$

$\langle$ BASIC COMMAND $\rangle \rightarrow \langle$ OPERATOR $\rangle$ $\langle$ THING SPEC $\rangle$

$\vdots$

E.G. FOR VEHICLE X IN DIVISION FOUR WHERE REPAIR-DATE OF X IS AFTER JANUARY 1973 AND VEHICLE-CODE OF X IS JEEP GIVE BREAKDOWN OF REPAIR-COSTS OF X BY MAINTENANCE-CODE

Attachment 3

CATEGORY IV

APPROXIMATIONS TO FLUENT NATURAL ENGLISH

> Give me a breakdown of the
> repair costs of all the
> division four vehicles that
> have been repaired since
> January
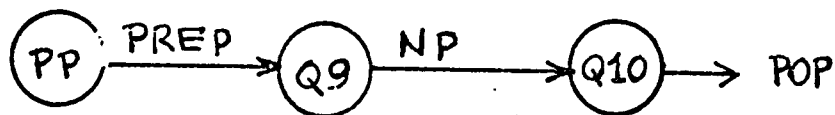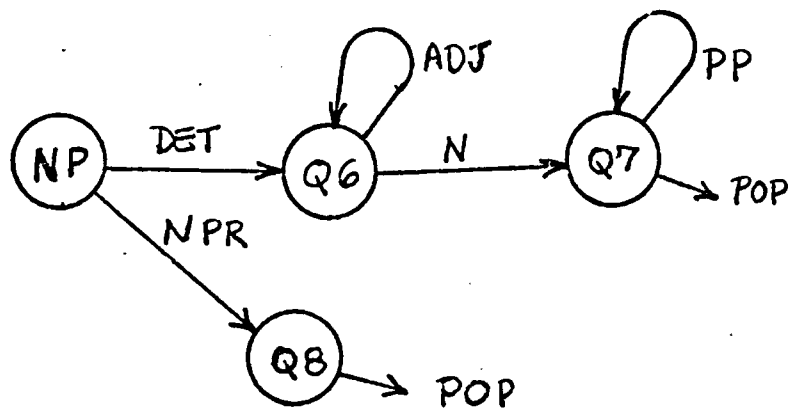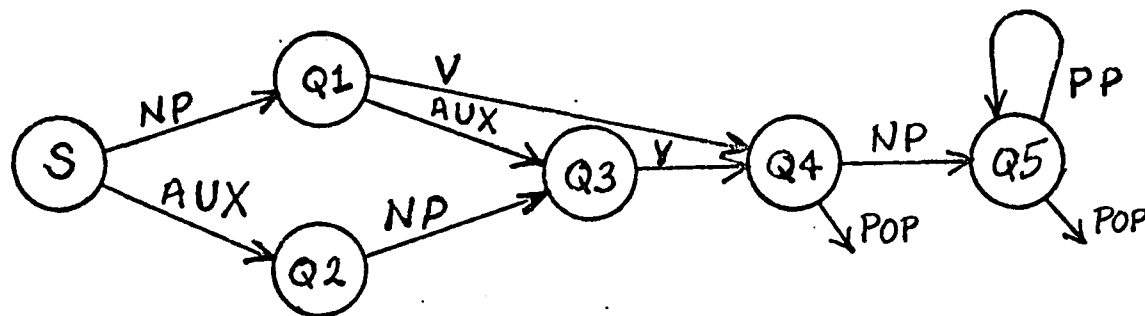
A SAMPLE TRANSITION NETWORK (BTN)

ONE LEVEL OF NETWORK FOR EACH TYPE
OF CONSTITUENT

Attachment 3

17

SMALL SEMANTIC NETWORK
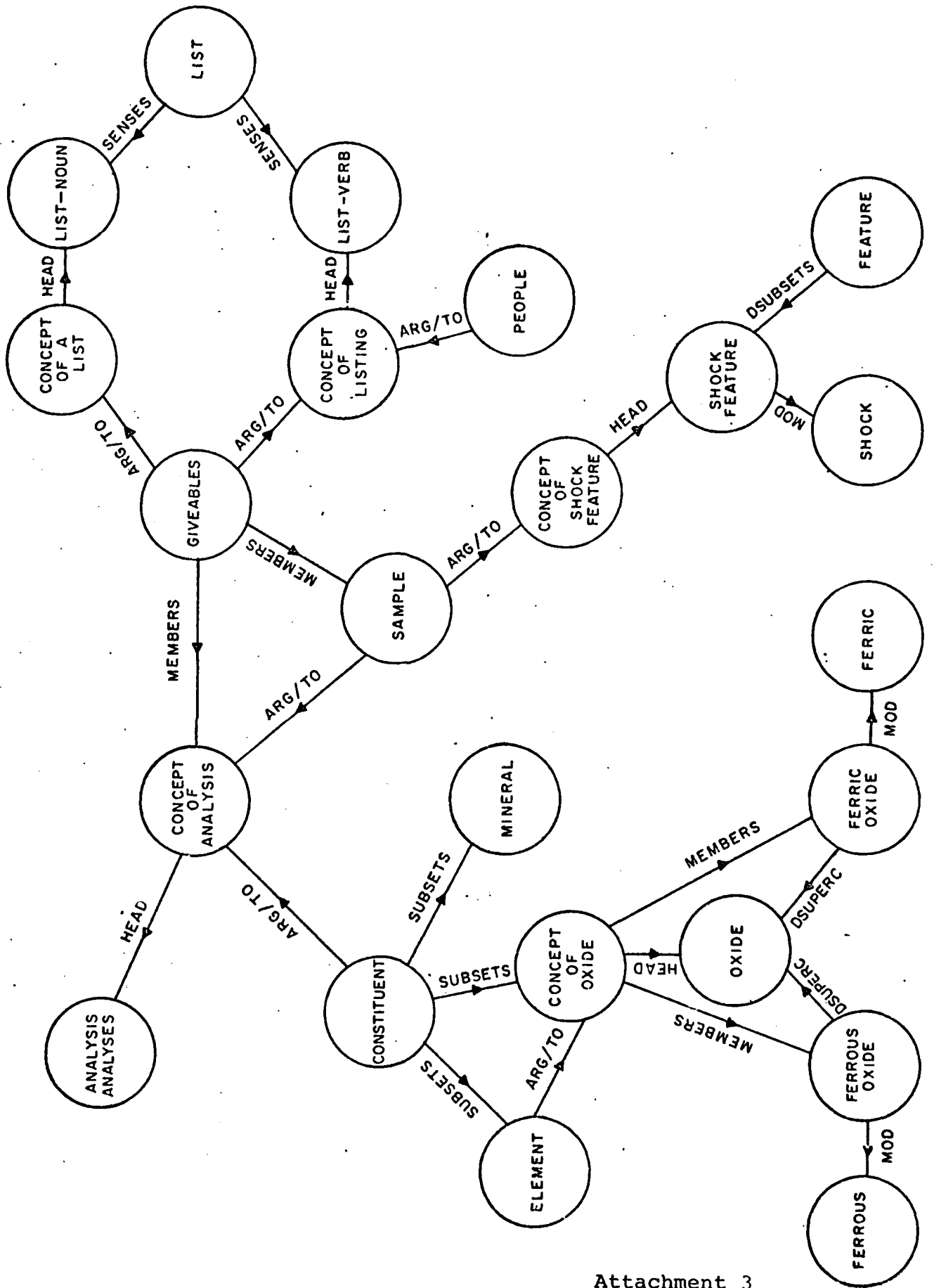
```
31**(HOW MANY BRECCIAS CONTAIN OLIVINE)
***
PARSING
815 CONSES
4.633 SECONDS
INTERPRETING
1574 CONSES
7.29 SECONDS
INTERPRETATIONS:
(FOR THE X12 / (SEQL (NUMBER X12 / (SEQ TYPECS) : (CONTAIN X12
(NPR* X14 / (QUOTE OLIV)) (QUOTE NIL)))) : T ; (PRINTOUT X12))

BBN LISP-10 03-09-72 ...
EXECUTING
(5)


-----


***
32**(WHAT ARE THEY)
***
PARSING
487 CONSES
2.755 SECONDS
INTERPRETING
1158 CONSES
4.053 SECONDS
INTERPRETATIONS:
(FOR EVERY X12 / (SEQ TYPECS) : (CONTAIN X12 (NPR* X14 / (QUOTE
OLIV)) (QUOTE NIL)) ; (PRINTOUT X12))

BBN LISP-10 03-09-72 ...
EXECUTING
S10059
S10059
S10065
S10067
S10073


-----
```

Attachment 3

INCREMENTAL SIMULATION

"IMPLEMENT" THE SYSTEM AS A COMBINATION
OF HUMAN SIMULATION AND COMPUTER PROGRAMS AND
"RUN" IT TO DISCOVER AND TEST ALGORITHMS AND
TO DEVELOP AN INTUITION FOR THE PROBLEM. THEN
GRADUALLY REPLACE THE HUMAN SIMULATORS WITH
COMPUTER CODE UNTIL THE HUMAN ROLE BECOMES ONE
OF EVALUATION AND TUNING.

Attachment 3

THE SDC/SRI SPEECH UNDERSTANDING SYSTEM

GOAL

DEVELOPMENT OF A SUS CAPABLE OF ENGAGING A HUMAN
OPERATOR IN A CONVERSATION ABOUT A SPECIFIC TASK
DOMAIN

TASK DOMAIN

DATA MANAGEMENT ON ATTRIBUTES OF WARSHIPS OF U.S.,
U.S.S.R., AND U.K.

PROJECT RESPONSIBILITIES

SDC	SRI
SIGNAL PROCESSING	SYNTAX
ACOUSTIC-PHONETICS	SEMANTICS
WORD & PHRASE PATTERN-MATCHING	PRAGMATICS
PROSODIC ANALYSIS	DISCOURSE ANALYSIS
SYSTEM HARDWARE & SOFTWARE	PARSING & SYSTEM CONTROL
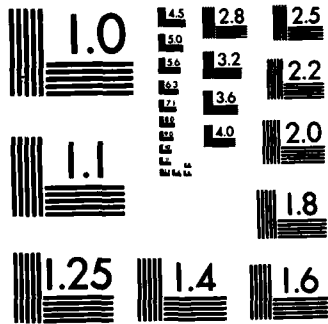
MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

DATA BASE

- CONTAINS (UNCLASSIFIED) INFORMATION ON
  265 WARSHIPS OF U.S., U.S.S.R. & U.K.

- EXTRACTED FROM JANE'S FIGHTING SHIPS

- DATA BASE EXPANSIONS AND UP-DATES
  COORDINATED WITH NELC

Attachment 4

2

```
                          ┌──────────┐
                          │ SPEAKER  │
                          └──────────┘
       ┌──────────────────┘          └──────────────────┐
┌────────────┐        ┌───────────┐         ┌────────────┐
│ ACOUSTIC   │ ─ ─ ─ ─│ EXECUTIVE │─────────│ UTTERANCE  │
│ PROCESSOR  │        └───────────┘         │ GENERATOR  │
└────────────┘              │               └────────────┘
┌────────────┐              │               ┌────────────┐
│ ACOUSTIC-  │              │               │ DISCOURSE  │
│ PHONETIC   │              │               │ HISTORY    │
│ DATA       │              │               └────────────┘
└────────────┘              │
         ┌──────────────────┴──────────────────┐
         │        LANGUAGE DEFINITION           │
         │  P    P    S    S    R    D          │
         │  H    H    Y    E    E    I          │
         │  O    O    N    M    S    S          │
         │  N    N    T    A    P    C          │
         │  E    O    A    N    O    O          │
         │  T    L    X    T    N    U          │
         │  I    O         I    S    R          │
         │  C    G         C    E    S          │
         │  S    Y         S         E          │
         └───┬──────────┬──────┬────┬───────────┘
             │          │      │    │
┌──────────┐ │     ┌─────────────────┐    ┌──────────┐
│ MAPPER   │─┘     │ SEMANTIC        │────│ DATA     │
└──────────┘       │ MODEL           │    │ BASE     │
                   └─────────────────┘    └──────────┘
```

Attachment 4

3
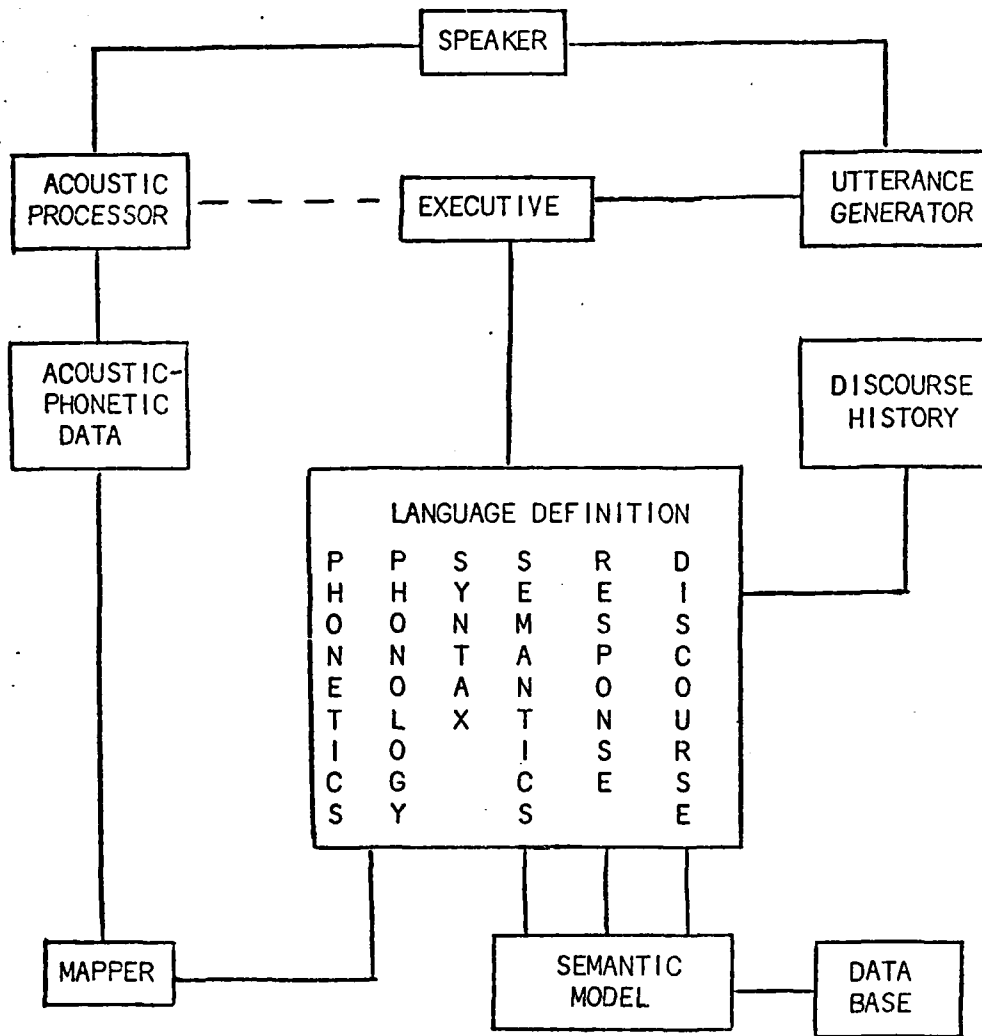
PARAMETRIZATION

- FUNDAMENTAL FREQUENCY EXTRACTION:

DECISION	ACCURACY
V/UV	99.4%
F_0	99.9%

- AUTOMATIC FORMANT FREQUENCY TRACKING

- SPECTRAL PARAMETERS FOR FRICATIVE/PLOSIVE ANALYSIS

Attachment 4

4

SEGMENTATION AND LABELING

VOWELS
&
SONORANTS

- FINDS SEGMENTS AND HANDLES COARTICULATION EFFECTS BY ANALYZING A FORMANT TRACK

- FINDS MULTIPLE LABELS WHERE APPROPRIATE AND REJECTS INAPPROPRIATE LABELS BY FORMANT PATTERN DISTANCE MEASURES

FRICATIVES
&
PLOSIVES

- PERFORMS MEANINGFUL ANALYSIS OF BRIEF AND RAPIDLY-CHANGING SOUNDS BY EMPLOYING PARAMETERS FROM NARROW-WINDOW SPECTRA

- HANDLES A WIDE VARIETY OF CLUSTER COARTICULATION EFFECTS

SYLLABLE
SEGMENTATION

- UTILIZES RESULTS OF PHONEMIC SEGMENTATION AND LABELING COMBINED WITH SPECIAL ENERGY FUNCTIONS TO LOCATE SYLLABLE BOUNDARIES

Attachment 4

WORD AND PHRASE PATTERN-MATCHING

- CONTAINS PHONOLOGICAL RULES COMPILER TO GENERATE ALTERNATE PRONUNCIATIONS OF LEXICAL BASE FORMS

- LEXICAL SUBSETTER EFFICIENTLY PRUNES LEXICON ON BASIS OF ACOUSTIC-PHONETIC DATA

- COMPACT REPRESENTATION OF PHONOLOGICAL VARIANTS PERMITS FAST MAPPING OF ALL RULE-GENERATED PRONUNCIATIONS

- DYNAMIC RATE-OF-SPEECH CALCULATIONS ACCURATELY RESTRICT LOCATIONS OF PREDICTED ITEMS IN ACOUSTIC-PHONETIC DATA WITHOUT RELYING ON PHONEME SEGMENTATION

Attachment 4

PROSODIC ANALYSIS

- ACOUSTIC PHRASES ARE SEGMENTED INTO SIMPLE
  CONTOURS OF FOUR TYPES: RISE-FALL, FALL-RISE,
  RISING, AND FALLING

- STRESS OF EACH SYLLABLE IS MARKED ON A FOUR-
  LEVEL SCALE BASED ON F_0, INTENSITY AND
  DURATION

- RATE-OF-SPEECH IS AUTOMATICALLY CALCULATED FROM
  SYLLABLE SEGMENTATION RESULTS

Attachment 4

7

EXECUTIVE

- PROVIDES OVERALL SYSTEM CONTROL

- COMBINES INPUTS FROM KNOWLEDGE SOURCES AT PHRASE LEVEL
  TO FORM OVERALL JUDGEMENT OF SUITABILITY OF INTERPRETATION

- WORKS TOP-DOWN (GOAL-DRIVEN) OR BOTTOM-UP (DATA-DRIVEN)
  BEGINNING AT ANY POINT IN THE UTTERANCE AND PROCESSING
  CONSTITUENTS IN EITHER DIRECTION

- ASSIGNS PRIORITIES TO TASKS IN TASK QUEUE BASED ON EXPECTED
  VALUES OF RESULTING INTERPRETATIONS AND CURRENT FOCUS
  OF ACTIVITY

- SHARES PARTIAL RESULTS AMONG COMPETING HYPOTHESES

Attachment 4

LANGUAGE DEFINITION

- INTEGRATES KNOWLEDGE SOURCES AT PHRASE LEVEL

- CONTAINS WORDS AND COMPOSITION RULES FOR COMBINING WORDS AND PHRASES USING ATTRIBUTE AND FACTOR STATEMENTS

- BASED ON PROTOCOLS FROM ACTUAL PERFORMANCE IN TASK-ORIENTED DIALOGS

- TUNEABLE TO DOMAIN OF DISCOURSE AND EASILY MODIFIED FOR DIFFERENT DOMAINS

- WRITTEN IN PERSON-ORIENTED LANGUAGE DEFINITION LANGUAGE AND COMPILED INTO COMPUTER-EFFICIENT FORM

Attachment 4

9

SEMANTICS

- CONTAINS NETWORK MODEL OF TASK DOMAIN AND COMPOSITION
  ROUTINES FOR COMBINING CONCEPTS OR FOR REJECTING
  MEANINGLESS COMBINATIONS

- PARTITIONS IN NETWORK FACILITATE ENCODING HIGHER-ORDER
  PREDICATES, MAINTAINING MULTIPLE PARSING HYPOTHESES, AND
  HANDLING QUANTIFICATION

- SHARES NETWORK SUBSTRUCTURES AMONG COMPETING HYPOTHESES

- MAINTAINS CORRESPONDENCE BETWEEN SYNTACTIC CATEGORY
  OF PHRASE AND NETWORK SUBSTRUCTURE

Attachment 4

DISCOURSE

- ENCODES DIALOG CONTEXT AND MAINTAINS DISCOURSE HISTORY

- EXPANDS ELLIPTICAL EXPRESSIONS BY PROCESSING NETWORK REPRESENTATIONS OF PRIOR UTTERANCES

- USES FOCUS SPACES TO IDENTIFY NOUN PHRASE REFERENTS AT. PHRASE LEVEL

Attachment 4

11

RESPONSE

- IDENTIFIES APPROPRIATE RESPONSE TO UTTERANCE BY SEARCHING DATA BASE

- SPECIFIES NETWORK SUBSTRUCTURE EMBODYING APPROPRIATE REPLY

Attachment 4

12

UTTERANCE GENERATOR

- CONVERTS NETWORK SUBSTRUCTURE INTO SENTENCE OR
  PHRASE FOR OUTPUT

Attachment 4

13

SYSTEM SOFTWARE

- NEW LIST PROCESSING LANGUAGE (CRISP) PROVIDES INCREASED ADDRESS SPACE AND EXPANDED ARITHMETIC CAPABILITIES

- INTERLISP/370 PROVIDES EXPANDED ADDRESS SPACE FOR IBM 370 LIST PROCESSING APPLICATIONS

- STANDARD DEC OPERATING SYSTEM (RSX-11M) USED FOR ACOUSTIC-PHONETIC PROCESSING.

Attachment 4

14

SYSTEM HARDWARE

SPEECH UNDERSTANDING SYSTEM TO BE IMPLEMENTED
ON THREE SDC COMPUTERS:

- IBM 370/145 FOR HIGHER-LEVEL LINGUISTIC
  PROCESSING AND WORD AND PHRASE PATTERN-
  MATCHING

- PDP-11/40 FOR SEGMENTATION AND LABELING

- SPS-41 FOR PARAMETRIZATION

Attachment 4

15

ATTACHMENT 5

FEATURES OF CMU SPEECH RESEARCH

1. GENERAL MODEL

2. MULTIPLE SYSTEMS

3. AUTOMATIC KNOWLEDGE AQUISION

4. PERFORMANCE ANALYSIS

5. THEORY

Attachment 5

1

FEATURES OF CMU SPEECH RESEARCH

1. <u>GENERAL MODEL</u>

   Explore Many Alternative Solutions to the
   Speech Understanding Problem

2. <u>MULTIPLE SYSTEMS</u>

	MANY KS	SYNTAX-DRIVEN
PDP-10	FASE OF EXPERIMENTATION	
C. mmp	16 PROCESSORS	
PDP-11/40 (Microcode)	LOW COST SUS	

3. <u>AUTOMATIC AND SEMI-AUTOMATIC KNOWLEDGE</u>

   AQUISION (LEARNING - PATTERNS OF SYMBOLS)

   ALLOPHONIC VARIABILITY
   COARTICULATION
   JUNCTURE RULES
   WORD PRONUNCIATION

4. <u>PERFORMANCE ANALYSIS</u>

   a. EXPLORE DESIGN CHOICES
   b. CLOSE TO REAL-TIME
   c. ITERATIVE DESIGN

5. <u>THEORY</u>

   LANGUAGE DESIGN, COMPLEXITY ANALYSIS, GRAMMATICAL
   INFERENCE

Attachment 5

2

SUS USING MANY DIFFERENT KNOWLEDGE SOURCES

(1972) HEARSAY - I ON CHESS TASK - TELEPHONE INPUT

     52%   SENTENCE ACCURACY WITHOUT SEMANTICS

     80%   SENTENCE ACCURACY WITH SEMANTICS

             (6 TIMES REAL TIME)

(1975) HEARSAY II ON NEWS RETRIEVAL TASK

          JUST STARTED WORKING

       15 DIFFERENT KNOWLEDGE SOURCES

SUS USING SYNTAX AND LEXICON

(1974) DRAGON ON FORMANT TASK (194 WORD VOC)

     31%   SENTENCE ACCURACY (122 TIMES REAL TIME)

     81%   WORD ACCURACY

(1975) HAPPY ON FORMANT TASK

     88%   SENTENCE ACCURACY (24 TIMES REAL TIME)

(FEW WEEKS) HAPPY ON PROG. LANGUAGE TASK

       3 SPEAKERS - HIGH BRANCHING FACTOR

SENTENCES		WORDS
80 TO 100%	ON TRAINING SENTENCES	(95-100%)
25 TO 48%	ON TEST SENTENCES	(75-84%)

PHRASE 12 IS THERE ANY NEWS ABOUT DEMOCRATS (383860/310184)

IS IS IS IS THERE THERE THERE THERE NEW

THERE THERE ANY ANY ANY ANY ANY NEWS NEWS NEWS NEWS

NEWS NEWS NEWS NEWS NEWS NEWS NEWS ABOUT ABOUT ABOUT ABOUT

ABOUT ABOUT ABOUT DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEM

DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS
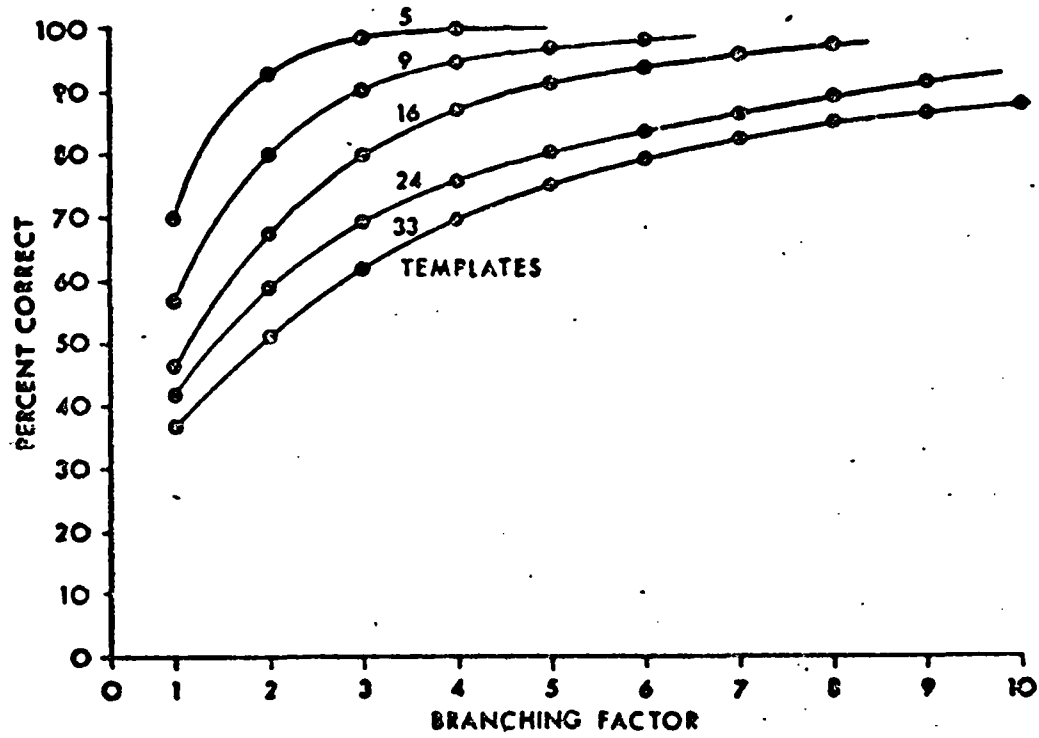
TO GET HIGH ACCURACIES

        CAREFUL TUNING OF THE SYSTEM IS ESSENTIAL

        REQUIRES MANY MANY RUNS ON TRAINING DATA

➡ CLOSE TO REAL TIME IS HIGHLY DESIRABLE

        100 SENTENCES OF 3 SECS - 5 MIN OF SPEECH

           25 TIMES REAL TIME   ➡   2 HRS/RUN

➡   250 TIMES REAL TIME   ➡   20 HRS/RUN

SYSTEMS WITH MANY GOOD IDEAS OFTEN FAIL BECAUSE OF A FEW

        WEAK LINKS IF THEY ARE SLOW

MANY ITERATIONS OF DESIGN CHOICES ARE NECESSARY TO

        GET RELIABLE SYSTEMS

Attachment 5

TASK	SIZE OF VOC.	LANGUAGE		CONFUSABILITY	
		ENTROPY	EQV. BRANCHING FACTOR	ENTROPY	EQV. BRANCHING FACTOR
DIGITS	10	3.32	10	0.24	1.18
ALPHABET	26	4.70	26	2.43	5.39
ALPHA-DIGIT	36	5.17	36	2.29	4.89
CHESS	31	2.87	7.30	1.73	3.32
LINCOLN	237	2.84	7.18		
EXTENDED	411	3.36	12.61		
IBM	250	2.872	7.32		
PROG. LANG. (NO SYNTAX)	37	5.21	37.00	1.92	3.78

EFFECTIVE VOCABULARIES USED BY VARIOUS SYSTEMS

Attachment 5

ATTACHMENT 6

Haskins Laboratories Speech Understanding Program

ACOUSTIC CUES IN NATURAL SPEECH: APPLICATIONS TO SPEECH RECOGNITION

1.   HUMAN ANALYSIS OF ACOUSTIC DATA

        Segmentation, Context Effects, Lexical
        Representation, Syntactic Boundaries,
        Stress Determination

2.   MACHINE STRATEGIES FOR DEALING WITH ACOUSTIC DATA

        Syllabic Units, Hierarchic vs. Parallel
        Feature Extraction, Constraints on Syllable
        Structure

3.   IMPROVED ANALYSIS TOOLS: DIGITAL PATTERN PLAYBACK

4.   RELATED RESEARCH: PERCEPTION AND PRODUCTION OF SPEECH

SPERRY+UNIVAC
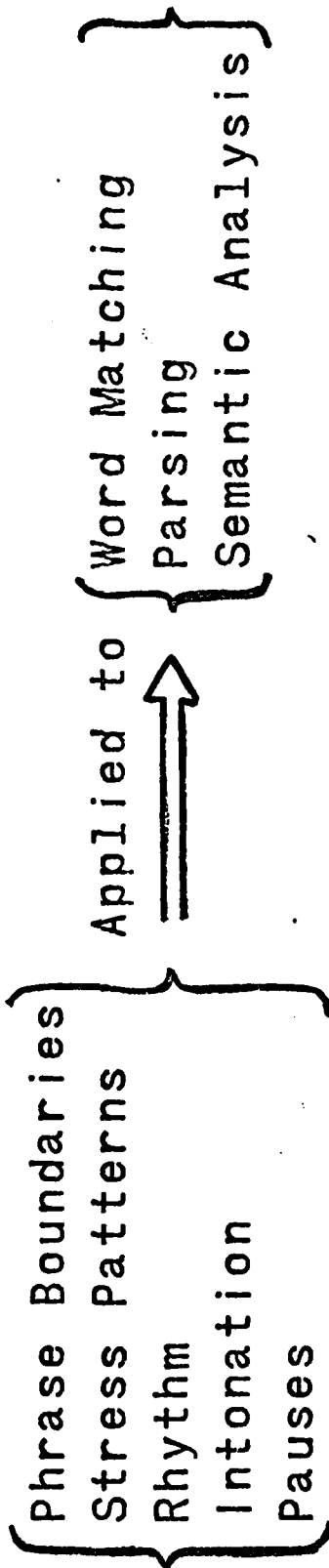
THE IMPORTANCE OF PROSODIC ANALYSES

IN

SPEECH UNDERSTANDING

- Reduce Computations

- Determine Sentence Type

- Disambiguate Sentence Structures

- Pauses Indicate Sentence and Clause Boundaries

- Stresses provide

  - Islands of phonetic reliability
  - Most Important Words
  - Phonetic/Phonemic Similarity
  - Subsetting of Lexicon
  - Conditions for Applying Phonological Rules
  - Rhythm and Rate of Speech
  - Syntactic Categories
  - Subordination Cues

SPERRY⬩UNIVAC

HIGHLIGHTS OF ARPA PROSODICS RESEARCH

EXPERIMENTS

- Machine classifications of phones work best in stressed syllables

- Listeners can consistently perceive which syllables are stressed

- Stressed syllables are accompanied by rising fundamental frequency and long high-energy syllabic nuclei

- Phrase boundaries are detectable from
  - Fall-rise valleys in fundamental frequency
  - Long intervals between stresses
  - Lengthened vowels and sonorant consonants

Attachment 7

3

SPERRY✦UNIVAC

HIGHLIGHTS OF ARPA PROSODICS RESEARCH

COMPUTER PROGRAMS

- Fundamental Frequency Tracking

- Detecting Phrase Boundaries from F_0

- Syllabification

- Stressed Syllable Locations

Attachment 7

4

SPERRY⊹UNIVAC

CONTRIBUTIONS OF ARPA RESEARCH

TO

SPERRY UNIVAC SPEECH PROJECTS

● Prosodically Guided Word Spotting System
  - Prosodic Guidelines
  - LPC Analysis
  - Phonetic Classifications
  - Segment Lattices
  - Word Matching and Scoring

● Speech Recognition Systems
  - Isolate Words
  - Connected Word Sequences

Attachment 7

5

ATTACHMENT 8

SCRL - ARPA PROJECT - NOVEMBER 13, 1975

OBJECTIVES:

- Accumulation of a large natural language data base transcribed orthographically, ARPAbetically (pseudo-phonemically), and phonetically.

- Development of computer programs for analyzing the above three levels of transcription.

- Editing the computerized natural language (SCRL) dictionary containing orthographic, phonemic, and gross grammatical codes.

- Compiling phonological rules for obtaining natural language variability from dictionary base form entries.

- Analysis of natural language data at the phonemic and phonetic levels.

- Support for ARPA Speech Understanding System (SUS) builders.

ACCOMPLISHMENTS:

- Data base: over 200 twenty-minute tape recordings obtained; over 30 of these transcribed orthographically and ARPAbetically; a few transcribed phonetically.

- Computer programs: complete set of programs to categorize, reference, and analyze the various levels of transcription, including cross-reference of data.

- Dictionary: editing completed and computer programs written for updating and maintaining the dictionary.

- Phonological rules: complete listing of rules from literature obtained; rules modified and new ones generated upon analysis of natural language data base.

- Analysis of data: statistical studies initiated on consonant clusters; vowel clusters; phonemic substitutions, deletions, and additions; phonemic frequency in various positions; phonemic-phonetic comparisons.

- System support: ARPAbetic transcriptions; base form dictionary entries; rule testing; Common Task Report.

Attachment 8

1

OBJECTIVES

1. Accumulation of a large natural language data base transcribed orthographically, ARPAbetically (pseudo-phonemically), and phonetically.

2. Development of computer programs for analyzing the above three levels of transcription.

3. Editing the computerized natural language (SCRL) dictionary containing orthographic, phonemic, and gross grammatical codes.

4. Compiling phonological rules for obtaining natural language variability from dictionary base form entries.

5. Analysis of natural language data at the phonemic and phonetic levels.

6. Support for ARPA Speech Understanding System (SUS) builders.

ACCOMPLISHMENTS

1. Data base: over 200 twenty-minute tape recordings obtained; over 30 of these transcribed ortho-graphically and ARPAbetically; a few transcribed phonetically.

2. Computer programs: complete set of programs to categorize, reference, and analyze the various levels of transcription, including cross-reference of data.

3. Dictionary: editing completed and computer programs written for updating and maintaining the dictionary.

4. Phonological rules: complete listing of rules from literature obtained; rules modified and new ones generated upon analysis of natural language data base.

5. Analysis of data: statistical studies initiated on consonant clusters; vowel clusters; phonemic sub-stitutions, deletions, and additions; phonemic frequency in various positions; phonemic-phonetic comparisons.

6. System support: ARPAbetic transcriptions; base form dictionary entries; rule testing; Common Task Report.

Attachment 8

ATTACHMENT 9

POTENTIAL SYSTEMS FOR MILITARY APPLICATIONS

1. <u>Digital Narrowband Communication System</u>: A massive effort is under way to develop and implement an all-digital communication system.

2. <u>Automatic Speaker Verification</u>: An advanced development model is being fabricated for secure access control applications.

3. <u>Training Systems</u>: A limited speech understanding system is under study for use as a component in a military training system.

4. <u>Distorted Speech Processing (Helium Speech)</u>: Helium speech unscramblers are being developed to allow for adequate diver-to-diver or diver-to-surface communications.

5. <u>On-Line Cartographic Processing System</u>: Studies are under way to use speech recognition and voice response techniques with cartographic point and trace processing systems.

6. <u>Word Recognition for Militarized Tactical Data Systems</u>: Word recognition, speaker verification and voice response will be used for message entry to a tactical data system.

7. <u>Voice Recognition and Synthesis for Aircraft Cockpit</u>: Existing word recognition systems are being tested and evaluated under simulated cockpit environments.

TABLE 1

MILITARY TASKS FOR POSSIBLE AUTOMATION

1. Security

    1.1 Speaker Verification (Authentication)

    1.2 Speaker Identification (Recognition)

    1.3 Determining emotional state of speaker (e.g.,
         stress effects)

    1.4 Recognition of spoken codes

    1.5 Secure access voice identification, whether or
         not in combination with fingerprints, facial
         information, identity card, signature, etc.

    1.6 Surveillance of communication channels

2. Command and Control

    2.1 System control (ships, aircraft, fire control,
         situation displays, etc.)

    2.2 Voice-operated computer input/output (each
         telephone a terminal)

    2.3 Data handling and record control

    2.4 Material handling (mail, baggage, publications,
         industrial applications)

    2.5 Remote control (dangerous material)

    2.6 Administrative record control

Attachment 9

TABLE 2

TECHNIQUES REQUIRING PERFECTING TO AUTOMATE MILITARY TASKS

1. Signal Conditioning:

Some processing of speech signals may be necessary to
compensate for different characteristics of input channels,
such as overall signal level and differential delay. Also,
it may be possible to preprocess to improve speech quality,
or S/N ratio, or to remove long silences.

2. Digital Signal Transformation:

The digitized speech signal is transformed in prepara-
tion for the extraction of parameters. Processes used
include Fourier and Walsh transforms, correlation, linear
predictive coding and digital filtering.

3. Analog Signal Transformation and Feature Extraction:

The signal can be transformed by hardware, such as
filter banks and correlation devices. Transforms can be
digitized for further processing or parameters and features
can be extracted in a continuous manner for presentation to
decision networks or algorithms.

4. Digital Parameter and Feature Extraction:

Calculations are done on the transformed signal to
extract relevant parameters, such as formant tracking,
pitch extraction and principle components analysis.

5.A Resynthesis:

Speech parameters extracted, as mentioned above, in
speech compression systems or stored in voice playback
systems must be retransformed into acceptable acoustic
speech signals.

Attachment 9

5.B Orthographic Synthesis:

In the translation of written materials to speech, a number of techniques must be developed. Some of these techniques are similar to those cited in the paragraphs above. One of the most important is the development of speech orthology.

6. Speaker Normalization, Speaker Adaptation, Situation Adaption:

The effectiveness of parameters in carrying relevant speech information depends on characteristics of individual speakers and on operational situations. This could mean that systems must be trained or must adapt to optimize parameters.

7. Time Normalization:

In recognition of isolated utterances, normalization is imposed to compensate for local and global differences in speech rate. Both linear and non-linear schemes can be used.

8. Segmentation and Labeling:

Segment boundaries are set, e.g., at points of rapid change, formant positions, voicing, spectral shape or other parameters. Segments may be labeled probabalistic-ally to acoustic-phonetic classes. Prestored knowledge of features and parameters for the various classes of segments are used in the decision.

9A. Language Statistics:

Language statistics and partial recognition are used to predict and evaluate words at specific points in an utterance.

9B. Syntax:

The grammar of the task is used to predict and evaluate word categories at specific points in an utterance.

9C. Semantics:

Knowledge of the task domain is used to predict and evaluate subject matter at specific points in an utterance.

9D. Speaker and Situation Pragmatics:

In determining the semantics of speech, certain aspects of the utterances are related to an underlying assumption about what the speaker would generally consider as an appropriate response. The development of this type of knowledge is required for speech understanding systems. Knowledge of the situation that gave rise to the speaker's utterances is also required for reliable interpretation and execution of the task to be performed in response to the utterance.

10. Lexical Matching:

Strings of linguistic-phonetic elements hypothesized by the linguistic part of the system are compared with strings of acoustic-phonetic elements derived from an utterance. A quantitative goodness of match is calculated.

11. Speech Understanding:

All sources of knowledge (acoustic, phonetic, pragmatic, semantic, syntactic) are used in combination to reconstruct the utterance and/or determine its meaning.

12. Speaker Recognition:

Speaker-specific parameters are extracted and compared with stored parameter sets from known speakers.

13.   System Organization and Realization:

Systems must be developed keeping in mind use by humans and cost-effective factors.

14.   Performance Evaluation:

Present development of all speech systems requires the determination of the quantitative value of each possible technique studied. Only by the use of stored speech samples is this performance evaluation possible.

Attachment 9

AUTOMATIC SPEECH PROCESSING REQUIREMENTS & COORD.

TAC	ROC	308-73	A THREE PART ENTRY/ACCESS CONTROL SYSTEM
SAC	ROC	1-71	AUTOMATIC PERSONNEL IDENTIFICATION & AUTHENTICATION SYSTEM
USAFSS	ROC	1-73	MODERNIZED AIRBORNE SIGINT SYSTEM
AFFDL	TN	12-72-33	EFFECTS OF STRESS ON HUMAN SPEECH
USAFSS	TWX	4-16-73	MESSAGE PRIORITIZING OPERATIONAL SPEECH PROCESSING RQMTS.

COORDINATION

USAFSS	ESD/BISS	TAC
SAC	USASA/LTC	AFAL
ARPA	AFOSR	AEC
NSA	CIA	FTD

SPEECH PROCESSING TECHNICAL COMMITTEES

IEEE	AUDIO & ELECTRO ACOUSTIC GROUP SPEECH PROCESSING COMMITTEE
IEEE	A & E SUBCOMMITTEE ON PATTERN RECOGNITION PROBLEMS IN SPEECH

Attachment 9

SPEAKER VERIFICATION FOR ENTRY CONTROL AT TI'S

CORPORATE INFORMATION CENTER (CIC)

MAJOR FEATURES:

⊗ RANDOMIZED SELECTION OF VERIFICATION PHRASES

GOOD	BEN	SWAM	NEAR
PROUD	BRUCE	CALLED	HARD
STRONG	JEAN	SERVED	HIGH
YOUNG	JOYCE	CAME	NORTH

⊙ VOICE PROMPTING

⊙ SEQUENTIAL DECISION STRATEGY

PERCENT OF USERS ACCEPTED:

ON 1ST PHRASE	76.7
2ND PHRASE	19.6
3RD PHRASE	2.9
4TH PHRASE	0.5
	(0.3)

1%

PERCENT OF USERS REJECTED:

PERCENT OF IMPOSTORS ACCEPTED:

ON 1ST PHRASE	0.3
2ND PHRASE	0.3
3RD PHRASE	0.2
4TH PHRASE	0.2
TOTAL:	(1.0)

2%

GRD 01-916   4/17/75

Attachment 9

AUTOMATIC
SPEAKER VERIFICATION PERFORMANCE

RADC PERFORMANCE (%)

BIAS MAXIMUM ERROR (%)	1 PHRASE	2 PHRASE	3 PHRASE	4 PHRASE
TRUE SPEAKER REJECTION (TYPE I ERROR)	1.0	1.0	1.0	1.0
IMPOSTOR ACCEPTANCE (TYPE 2 ERROR)	2.5	1.0	.03	0.0
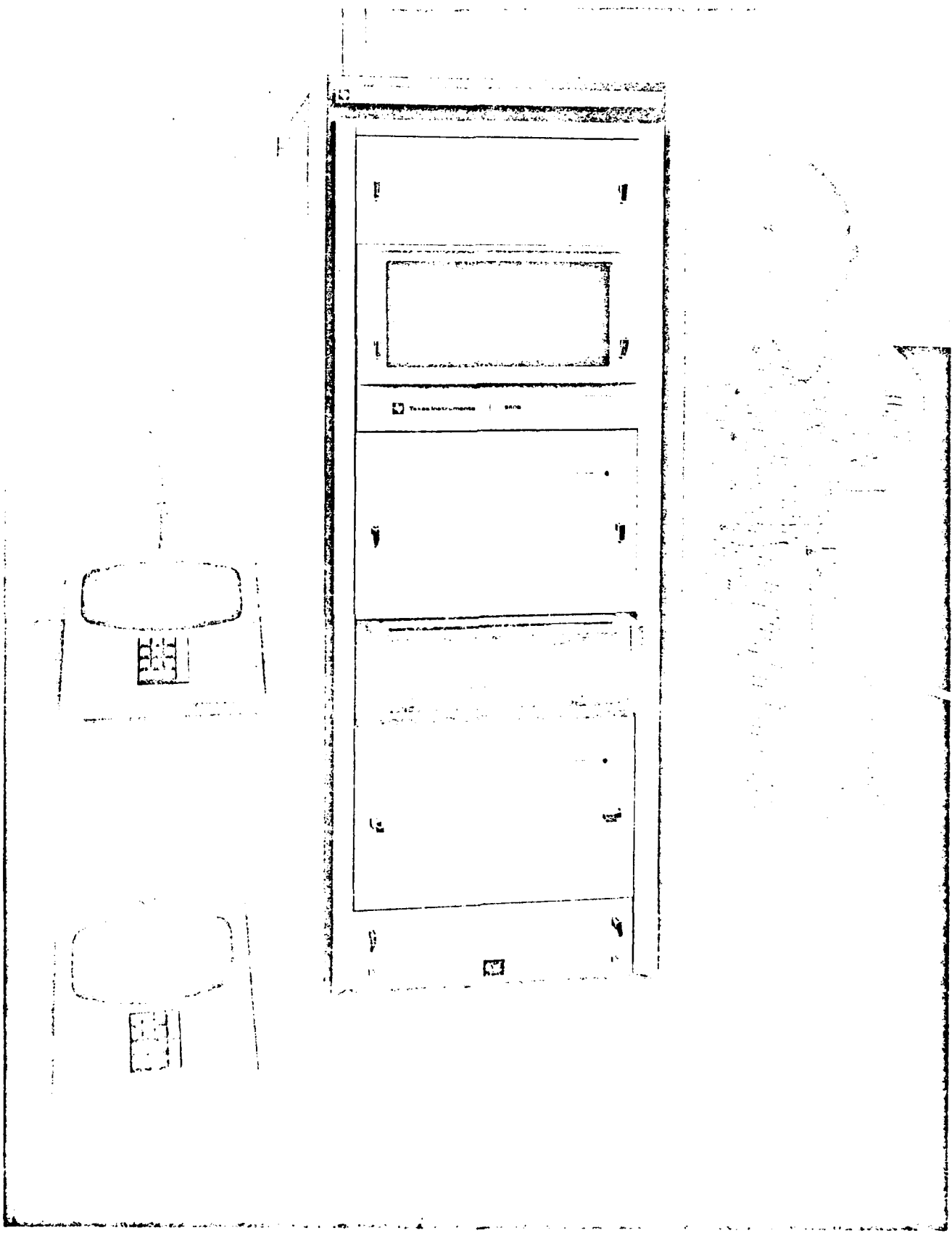
TEXAS INSTRUMENTS
INCORPORATED

DMA VOICE ENTRY SYSTEM SPECIFICATIONS

SELF CONTAINED/SELF OPERATING

VOCABULARY
- 10 DIGITS PLUS 5 CONTROL WORDS

TRAINING
- ON-LINE/REAL-TIME
- LESS THAN 10 SECONDS PER WORD

OPERATION
- IMMEDIATE RESPONSE TIME
- CAPABLE OF STORING MULTISPEAKER TRAINING DATA

OUTPUT
- VISUAL DISPLAY FOR OPERATOR VERIFICATION
- DIGITAL ENCODED OUTPUT
- HARD COPY TELETYPE PRINTOUT

PHYSICAL
- BASIC HARDWARE
- NOVA-1200 COMPUTER
- HIGH-SPEED PAPER TAPE READER
- STD ASR 33 TELETYPE
- CUSTOMIZED VISUAL DISPLAY

Attachment 9

KEYWORD DETECTION

INTERCEPT RECEIVER 1

INTERCEPT RECEIVER 2

INTERCEPT RECEIVER 3

INTERCEPT RECEIVER N

KEYWORD DETECTORS

1 2 3 · · · N

SWITCHING MATRIX

CONSOLE 1

CONSOLE 2

CONSOLE M

LINGUIST

LINGUIST

LINGUIST

N>M

Attachment 9
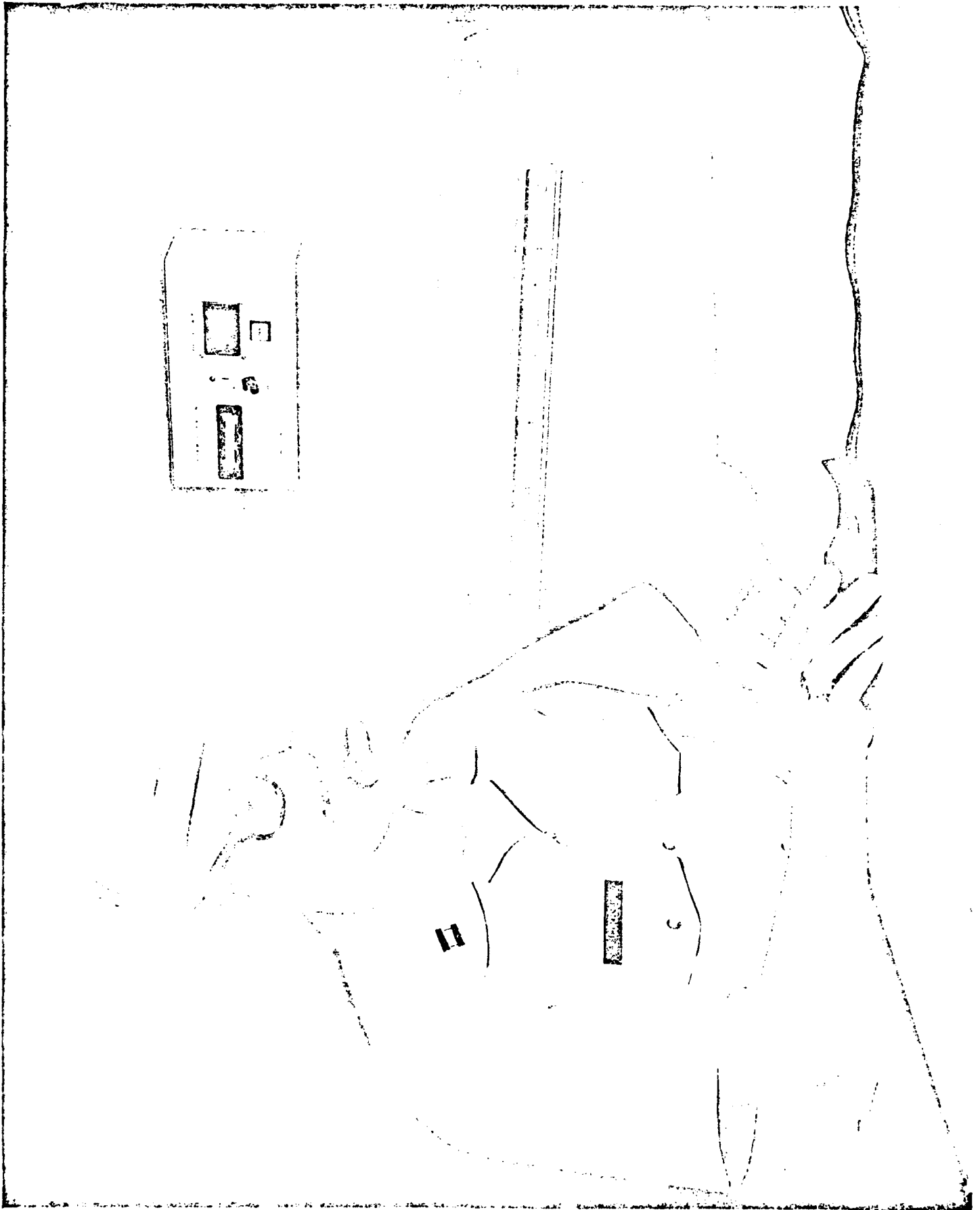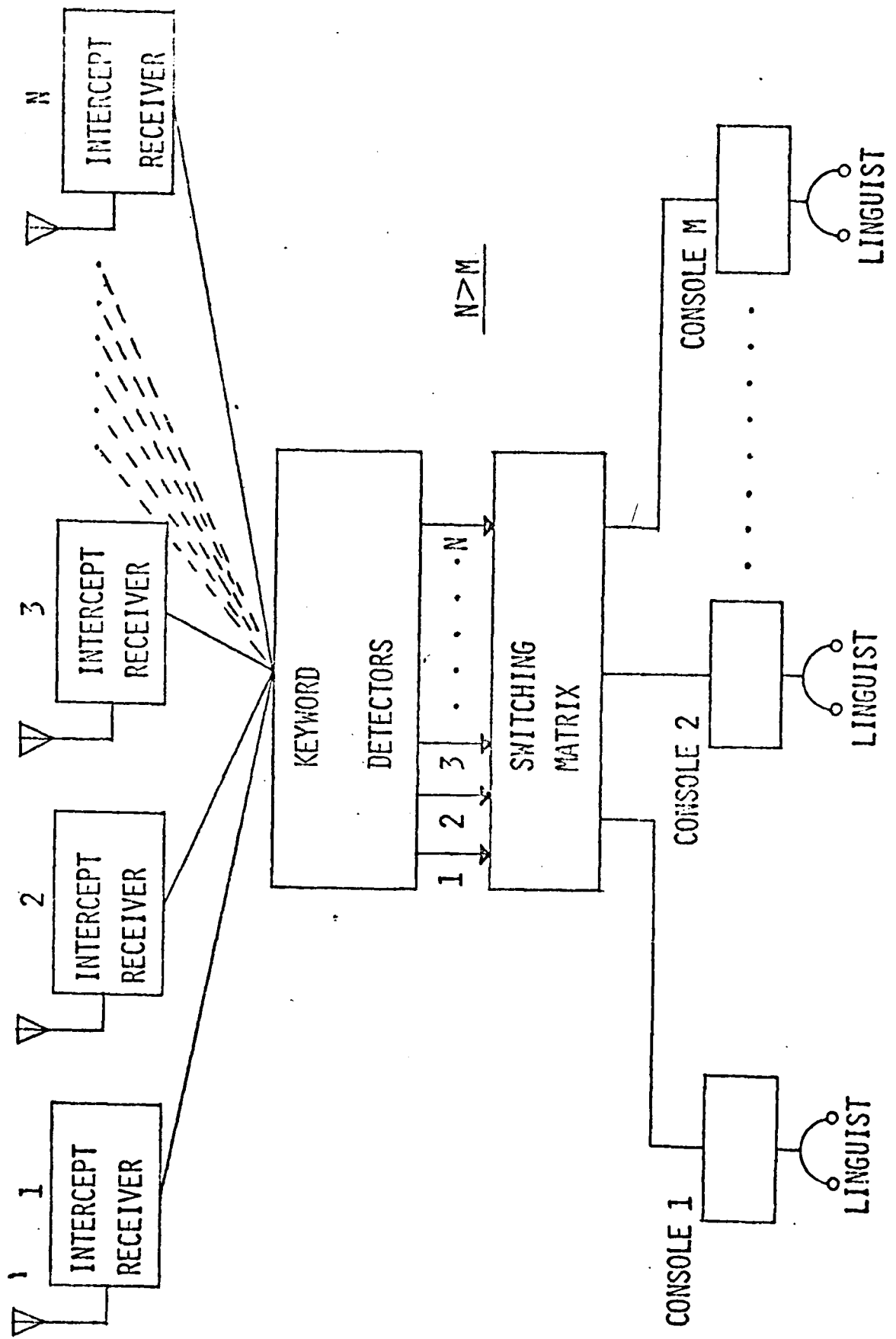
SPEECH ENHANCEMENT GOALS

- RECOGNITION OF VOICE MODULATED SIGNALS

- ENHANCEMENT OF SPEECH SIGNAL-TO-NOISE RATIO

- PREPROCESSOR AND NORMALIZER FOR AUTOMATIC RECOGNITION EQUIPMENT

RADC-IR-74-249

Attachment 9

AUTOMATIC SPEECH RECOGNITION

SECURE ACCESS
SPEAKER VERIFICATION

CLASSIFIED AREA

AUTHORIZED PERSONNEL ONLY
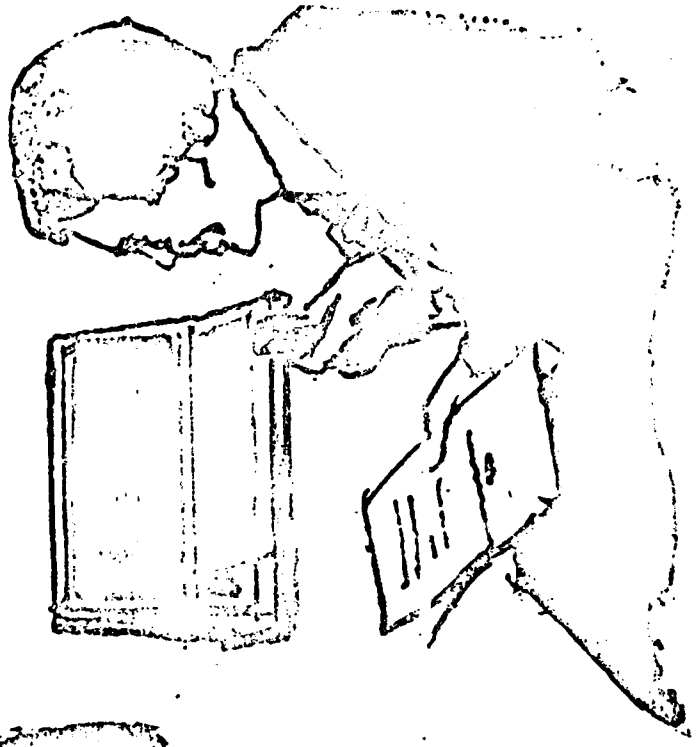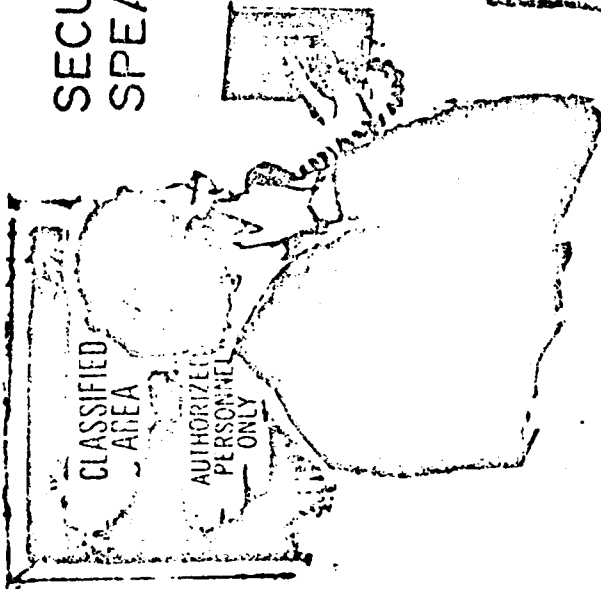
VOICE INPUT
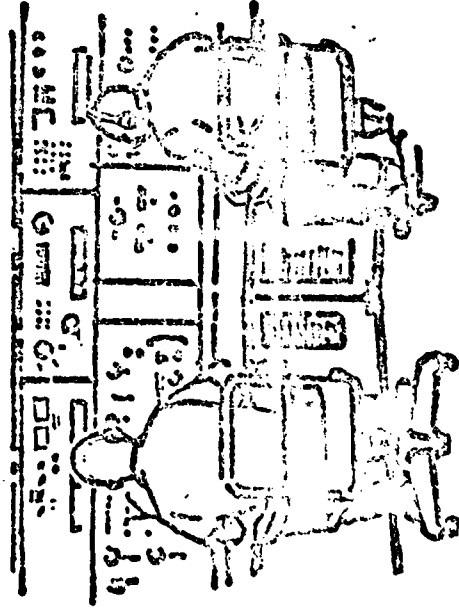
AUTOMATIC MESSAGE MONITORING

Attachment 9

NATO RSG-4 ASSESSMENT OF AUTOMATIC SPEECH RECOGNITION

David C. Hodge
Principal US Delegate
US Army Human Engineering Laboratory
Aberdeen Proving Ground, Maryland

Research Study Group 4 on Automatic Pattern Recognition is organized under the Defense Research Group (AC/243 Exploratory Panel III on Physics and Electronics. The function of RSG-4 is to assess military applications of automatic pattern recognition technology, to summarize the state of the art and to identify bases for cooperative research among the NATO countries.

Slide 1 lists the participating countries and the technical objectives of RSG-4.

Slide 2 lists the activities of RSG-4 in pursuing its technical objectives, the starting dates, and the present status.

Slide 3 outlines the steps taken in performing a technology assessment of the topic: "Automatic Speech Recognition for Military Applications."

Slide 4 lists the various military tasks (potential applications) in the speech area that we would like to be able to automate.

Slide 5 (3 pages) lists, and defines, the speech processing techniques that have to be perfected in order to be able to automate the military tasks.

Slide 6 gives the state of the art for each of the processing techniques, and lists the identification numbers of the unsolved problems for each technique.

Slide 7 presents a glossary of unsolved problems and their definitions.

Slide 8 lists near-term applications of speech processing and speech recognition technology to military problems. These applications are expected to be realized within the next decade or less (considerably less in some cases).

Attachment 10

AC/243 (PANEL III) RSG-4

TOPIC: AUTOMATIC PATTERN RECOGNITION

DATE CONSTITUTED: NOVEMBER 1971

PARTICIPATING COUNTRIES: CANADA, DENMARK, FRANCE, GERMANY, THE NETHERLANDS,

UNITED KINGDOM, UNITED STATES

TECHNICAL OBJECTIVES:

1. TO REVIEW APR TOPICS OF MILITARY SIGNIFICANCE AND RECOMMEND COOPERATIVE
   RESEARCH ACTIVITIES. (NOTE: NATO RSG'S PROVIDE A MECHANISM FOR
   INTERNATIONAL COOPERATION ON TOPICS FOR WHICH OTHER MECHANISMS
   ARE INAPPROPRIATE, E.G., CLASSIFIED APPLICATIONS.)

2. TO CONTINUE EXCHANGING INFORMATION ABOUT APR PROJECTS IN THE PARTICIPATING
   COUNTRIES.

Attachment 10

2

AC/243 (PANEL III) RSG-4

ACTIVITIES:

STATUS:

1. DEFINE CATEGORIES FOR CLASSIFYING APR RESEARCH PROJECTS — COMPLETED MAY 72

2. EXCHANGE SUMMARY REPORTS ON NATIONAL PROJECTS — 1972, 1973 **

3. IDENTIFY COMMON AREAS OF INTEREST (AND DISINTEREST) — CONTINUING

4. CONDUCT TECHNOLOGY ASSESSMENTS OF SELECTED TOPICS:

   A. AUTOMATIC IMAGE PROCESSING — STARTED NOV 73

   B. AUTOMATIC SPEECH RECOGNITION — STARTED MAY 74

   C. MECHANICAL WAVE PROCESSING TECHNIQUES — SCHEDULED FEB 76

5. DEVELOP COOPERATIVE RESEARCH PROPOSALS:

   A. IMAGE PROCESSING — PROPOSED AUG 74 / INITIATED JUL 75

(**SUMMARIES AVAILABLE: 1972, AD 905 384 L; 1973, AD 915 008 L.)

Attachment 10

3

AC/243 (PANEL III) RSG-4

OUTLINE OF TECHNOLOGY ASSESSMENT OF AUTOMATIC SPEECH RECOGNITION AND PROCESSING

1. NOVEMBER 1973: DEVELOPED LIST OF POSSIBLE (PROBABLE) MILITARY APPLICATIONS IN THIS AREA.

2. MAY 1974: LIST OF APPLICATIONS REVISED. SPECIALISTS FROM PARTICIPATING COUNTRIES PRESENTED INDEPENDENT ASSESSMENTS OF (A) THE STATE OF THE ART, (B) UNSOLVED PROBLEMS, (C) ESTIMATED COST OF SOLUTION, AND (D) PROBABLE SYSTEM REQUIREMENTS FOR REALIZATION. UNITED STATES ASSESSMENT PREPARED BY NED NEUBURG (USA) WITH INPUTS FROM B. BEEK, R. VONUSA, T. MCDONALD, R. BOUVIER (RADC), J. DECLERK (USAECOM), R. KAHN (DARPA), T. TREMAINE (USA), D. HODGE (USAHEL).

3. AUGUST 1974: FRANCE, GERMANY, THE NETHERLANDS, AND UNITED STATES EXPRESSED SOME DEGREE OF INTEREST IN COOPERATION ON ONE OR MORE ASPECTS. (NO DEFENSE-SUPPORTED WORK IN CANADA, DENMARK OR UNITED KINGDOM.) PRIMARY INTEREST WAS EXPRESSED BY THE NETHERLANDS AND UNITED STATES.

4. JULY 1975: DRAFT TECHNOLOGY SUMMARY REPORT, PREPARED BY BRUNO BEEK (RADC) REVIEWED BY RSG-4. FINAL FORMAT DISCUSSED. NO FURTHER BASIS FOR COOPERATION FOUND.

5. NOVEMBER 1975: FINAL VERSION OF TECHNOLOGY ASSESSMENT TO BE SUBMITTED FOR PUBLICATION AS NATO WORKING PAPER.

6. FEBRUARY 1976: DISCUSSION OF POSSIBLE COOPERATION ON CLASSIFIED APPLICATIONS OF SPEECH PROCESSING AND SPEECH RECOGNITION. IF INSUFFICIENT INTEREST IS EXPRESSED, TOPIC WILL BE DROPPED.

Attachment 10

AC/243 (Panel III) RSG-4

Military Tasks for Possible Automation

1. Security Applications
    A. Speaker verification
    B. Speaker identification
    C. Speaker discrimination

2. Data Transmission and Communication
    A. Conventional vocoders
    B. LPC vocoders
    C. Other vocoders
    D. Adapting systems to voice changes

3. Voice-Operated Systems

    A. Limited Word Sets
        (1) Isolated words
        (2) Short phrases

    B. Continuous Speech
        (1) Keyword spotting
        (2) Speech understanding

4. Distorted Speech (Helium speech)

5. Language Identification

Attachment 10

TABLE 2

TECHNIQUES REQUIRING PERFECTING TO AUTOMATE MILITARY TASKS

1.   Signal Conditioning:

Some processing of speech signals may be necessary to
compensate for different characteristics of input channels,
such as overall signal level and differential delay. Also,
it may be possible to preprocess to improve speech quality,
or S/N ratio, or to remove long silences.

2.   Digital Signal Transformation:

The digitized speech signal is transformed in prepara-
tion for the extraction of parameters. Processes used
include Fourier and Walsh transforms, correlation, linear
predictive coding and digital filtering.

3.   Analog Signal Transformation and Feature Extraction:

The signal can be transformed by hardware, such as
filter banks and correlation devices. Transforms can be
digitized for further processing or parameters and features
can be extracted in a continuous manner for presentation to
decision networks or algorithms.

4.   Digital Parameter and Feature Extraction:

Calculations are done on the transformed signal to
extract relevant parameters, such as formant tracking,
pitch extraction and principle components analysis.

5.A Resynthesis:

Speech parameters extracted, as mentioned above, in
speech compression systems or stored in voice playback
systems must be retransformed into acceptable acoustic
speech signals.

Attachment 10

6

5.B Orthographic Synthesis:

In the translation of written materials to speech, a number of techniques must be developed. Some of these techniques are similar to those cited in the paragraphs above. One of the most important is the development of speech orthology.

6. Speaker Normalization, Speaker Adaptation, Situation Adaption:

The effectiveness of parameters in carrying relevant speech information depends on characteristics of individual speakers and on operational situations. This could mean that systems must be trained or must adapt to optimize parameters.

7. Time Normalization:

In recognition of isolated utterances, normalization is imposed to compensate for local and global differences in speech rate. Both linear and non-linear schemes can be used.

8. Segmentation and Labeling:

Segment boundaries are set, e.g., at points of rapid change, formant positions, voicing, spectral shape or other parameters. Segments may be labeled probabalistic- ally to acoustic-phonetic classes. Prestored knowledge of features and parameters for the various classes of segments are used in the decision.

9A. Language Statistics:

Language statistics and partial recognition are used to predict and evaluate words at specific points in an utterance.

Attachment 10

9B. Syntax:

The grammar of the task is used to predict and evaluate word categories at specific points in an utterance.

9C. Semantics:

Knowledge of the task domain is used to predict and evaluate subject matter at specific points in an utterance.

9D. Speaker and Situation Pragmatics:

In determining the semantics of speech, certain aspects of the utterances are related to an underlying assumption about what the speaker would generally consider as an appropriate response. The development of this type of knowledge is required for speech understanding systems. Knowledge of the situation that gave rise to the speaker's utterances is also required for reliable interpretation and execution of the task to be performed in response to the utterance.

10. Lexical Matching:

Strings of linguistic-phonetic elements hypothesized by the linguistic part of the system are compared with strings of acoustic-phonetic elements derived from an utterance. A quantitative goodness of match is calculated.

11. Speech Understanding:

All sources of knowledge (acoustic, phonetic, pragmatic, semantic, syntactic) are used in combination to reconstruct the utterance and/or determine its meaning.

12. Speaker Recognition:

Speaker-specific parameters are extracted and compared with stored parameter sets from known speakers.

Attachment 10

13. System Organization and Realization:

Systems must be developed keeping in mind use by humans and cost-effective factors.

14. Performance Evaluation:

Present development of all speech systems requires the determination of the quantitative value of each possible technique studied. Only by the use of stored speech samples is this performance evaluation possible.

Attachment 10

TABLE 3   SLIDE 6

STATE OF THE ART AND UNSOLVED PROBLEMS

PROCESSING TECHNIQUES	STATE OF THE ART[1]	UNSOLVED PROBLEMS[2]
1. Signal Conditioning	A, except speech enhancement (C)	1, 15, 20, 23
2. Digital Signal Transformation	A	1, 15, 20
3. Analog Signal Transformation and Feature Extraction	A, except feature extraction (C)	1, 2, 6, 14-16, 20 24, 25
4. Digital Parameter and Feature Extraction	B	1, 2, 6, 14, 16, 20, 24, 25
5A Resynthesis	A	4, 7, 20, 25
5B Orthographic Synthesis	C	4, 6-8, 19, 26-28 29
6. Speaker Normalization, Speaker Adaptation Situation Adaptation	C	15-17, 19, 20, 23, 24, 25, 29
7. Time Normalization	B	3, 16, 20, 25, 29
8. Segmentation and Labeling	B	1, 4, 5-9, 11, 13, 16, 18-20, 24, 25
9A Language Statistics	C	5, 8, 9, 11, 12, 14, 20, 24, 25
9B Syntax	B	6, 7, 9, 12, 14, 20, 25
9C Semantics	C	6, 7, 9, 10, 12, 14, 20, 25
9D Speaker and Situation Pragmatics	C	3, 6, 12, 14, 16, 18, 19, 23
10 Lexical Matching	C	7-9, 12-14, 20, 25
11 Speech Understanding	B-C	5, 9, 12, 14, 16, 18, 23, 25
12 Speaker Recognition	A for speaker verification; c for all others	14, 16, 17, 19, 20, 24, 25
13 System Organization and Realization	A-C	21, 22
14 Performance Evaluation	C	1, 6-11, 18-20, 24-28

[1] Ratings: A = Useful Now; B = Shows Promise; C = A Long Way to Go

[2] See Glossary (Table 4) for problem definitions; list may not be exhaustive.

TABLE 4

GLOSSARY OF PROBLEMS AND DEFINITIONS

1. Detect speech in noise; speech/non-speech.

2. Extract relevant acoustic parameters (poles, zeros, formant (transitions), slopes, dimensional representation, zero-crossing distributions).

3. Dynamic programming (non-linear time normalization).

4. Detect smaller units in continuous speech (word/phoneme boundaries; acoustic segments).

5. Establish anchor point; scan utterance from left to right; start from stressed vowel; etc.

6. Stressed/unstressed.

7. Phonological rules.

8. Missing or extra added ("uh") speech sound.

9. Limited vocabulary and restricted language structure necessary; possibility of adding new words.

10. Semantics of (limited) tasks.

11. Limits of acoustic information only.

12. Combine acoustical, syntax and semantic information.

13. Recognition algorithm (shortest distance, (pairwise) discriminant, Bayes, probabilities).

14. Hypothesize-and-test, backtrack, feed forward.

15. Effect of nasalization, cold, emotion, loudness, pitch, whispering, distortions due to talker's acoustical environment, distortions by communication systems (telephone, transmitter-receiver, intercom, public address, face masks), non-standard environments.

16. Adaptive and interactive quick learning.

17. Mimicking; uncooperative speaker(s).

18. Necessity of visual feedback, error control, level for rejections.

19. Constancy of references.

20. Real-time processing.

21. Human engineering problem of incorporating speech understanding system into actual situations.

22. Cost effectiveness.

23. Detect speech in presence of competing speech.

Attachment 10

SLIDE 7

Glossary (Cont'd)

24. Economical ways of adding new speakers to system.
25. Use of prosodic information.
26. Coarticulation rules.
27. Morphology rules.
28. Syntax rules.
29. Vocal tract modeling.

Attachment 10

AC/243 (PANEL III) RSG-4

NEAR-TERM APPLICATIONS OF SPEECH RECOGNITION & PROCESSING TO MILITARY PROBLEMS

1. DIGITAL NARROWBAND COMMUNICATION SYSTEM: A MASSIVE EFFORT IS UNDER WAY TO DEVELOP AND IMPLEMENT AN ALL-DIGITAL COMMUNICATION SYSTEM.

2. AUTOMATIC SPEAKER VERIFICATION: AN ADVANCED DEVELOPMENT MODEL HAS BEEN FABRICATED FOR SECURE ACCESS CONTROL APPLICATIONS.

3. TRAINING SYSTEMS: A LIMITED SPEECH UNDERSTANDING SYSTEM IS UNDER STUDY FOR USE AS A COMPONENT IN A MILITARY TRAINING SYSTEM.

4. DISTORTED SPEECH PROCESSING (HELIUM SPEECH): HELIUM SPEECH UNSCRAMBLERS ARE BEING DEVELOPED TO ALLOW ADEQUATE DIVER-TO-DIVER AND DIVER-TO-SURFACE COMMUNICATIONS.

5. ON-LINE CARTOGRAPHIC PROCESSING SYSTEM: STUDIES ARE UNDER WAY TO USE SPEECH RECOGNITION AND VOICE RESPONSE TECHNIQUES WITH CARTOGRAPHIC POINT AND TRACE PROCESSING SYSTEMS.

6. WORD RECOGNITION FOR MILITARIZED TACTICAL DATA SYSTEMS: WORD RECOGNITION, SPEAKER VERIFI-CATION AND VOICE RESPONSE WILL BE USED FOR MESSAGE ENTRY INTO A TACTICAL DATA SYSTEM.

7. VOICE RECOGNITION AND SYNTHESIS FOR AIRCRAFT COCKPIT: EXISTING WORD RECOGNITION SYSTEMS ARE BEING TESTED AND EVALUATED UNDER SIMULATED COCKPIT ENVIRONMENTS.

Attachment 10

DoD Speech Environment

- Large Volumes of Speech

- Many Talkers

- Limited Bandwidth

- Often Low S/N

- Varied Semantic Environment

- Poor Microphones

- Real Time Processing

Speech Research Interests

(I) Automate Speech Processing

- Words (Meaning?)
- Language Spoken
- Talker Identity

(II) Efficient Speech Coding

- Storage Costs
- Transmission Costs

(III) Intelligibility Enhancement

- Suppress Noise Effects

Attachment 11

2

	SPEECH UNDERSTANDING SYSTEM	KEYWORD SPOTTING SYSTEM
1)	ACCEPT CONTINUOUS SPEECH	ACCEPT CONTINUOUS SPEECH
2)	FROM MANY	FROM AN UNLIMITED SET OF
3)	COOPERATIVE TALKERS OF THE GENERAL AMERICAN DIALECT	"NOT UNCOOPERATIVE" TALKERS OF SEVERAL LANGUAGES
4)	IN A QUIET ROOM	IN A NOISY ENVIRONMENT
5)	OVER A GOOD QUALITY MICROPHONE	GIVEN TELEPHONE BANDWIDTH
6)	ALLOWING SLIGHT TUNING OF THE SYSTEM PER SPEAKER	AUTOMATICALLY ADJUSTING TO THE TALKER
7)	REQUIRING ONLY NATURAL ADAPTATION BY THE USER	WITH NO ADATATION BY THE SPEAKER
8)	PERMITTING A SLIGHTLY SELECTED VOCABULARY OF 1000 WORDS	PERMITTING A SLIGHTLY SELECTED AND READILY CHANGEABLE VOCABULARY
9)	WITH A HIGHLY ARTIFICIAL SYNTAX	WITH NO SYNTACTIC SUPPORT
10)	ON A SPECIFIED TASK	IN AN UNSPECIFIED TASK ENVIRONMENT
11)	WITH A SIMPLE PSYCHOLOGICAL MODEL OF THE USER	WITH AN INADEQUATE PSYCHOLOGICAL MODEL OF THE USER
12)	PROVIDING GRACEFUL INTERACTION	AND NO INTERACTION
13)	TOLERATING LESS THAN 10 % SEMANTIC ERROR	TOLERATING LESS THAN 1 % FALSE ALARMS WITH LESS THAN 30 % FALSE DISMISSALS
14)	IN A FEW TIMES REAL TIME	IN AT MOST REAL TIME
19)	DEMONSTRABLE IN 1976 WITH A MODERATE CHANCE OF SUCCESS	DEMONSTRABLE IN 1978

J.F.BOEHM   R543

Attachment 11

3

Word Recognition Research Emphasis

Develop System:

- Words in Natural Context
- Normalize Talker Differences
- Recognize Talker Change
- Process Telephone Quality Speech
- Normalize Channel Distortion
- Apply Phonological Constraints
- Easy Lexical Entry

Study:

- Phonetic Shifts (Dialects)
- Impact of Language Structure
- Probability of Errors (I & II)
- Limitations of No Syntax/Semantics

Attachment 11

4

INCOHERENT

ELECTROOPTICAL PROCESSING

WITH CHARGE

TRANSFER DEVICES

Attachment 12

BASIC COMPONENTS

o Light Emitting Diodes
     Temporal Modulation of Light

o Optical Transparencies
     Spatial Modulation of Light

o Charge Coupled Devices
     Integration
     Shift Registration

Attachment 12

AREA ARRAY PROCESSOR

Light-Emitting Diode

Condensing Lens

Optical Transparency

Imaging Lens

Area Charge Coupled Device

NELC SERIES

Attachment 12

3

ELECTRO-OPTICAL PROCESSING

$$f(u) \longrightarrow \boxed{h(v,u)} \longrightarrow g(v)$$

INPUT

IMPULSE RESPONSE

OUTPUT
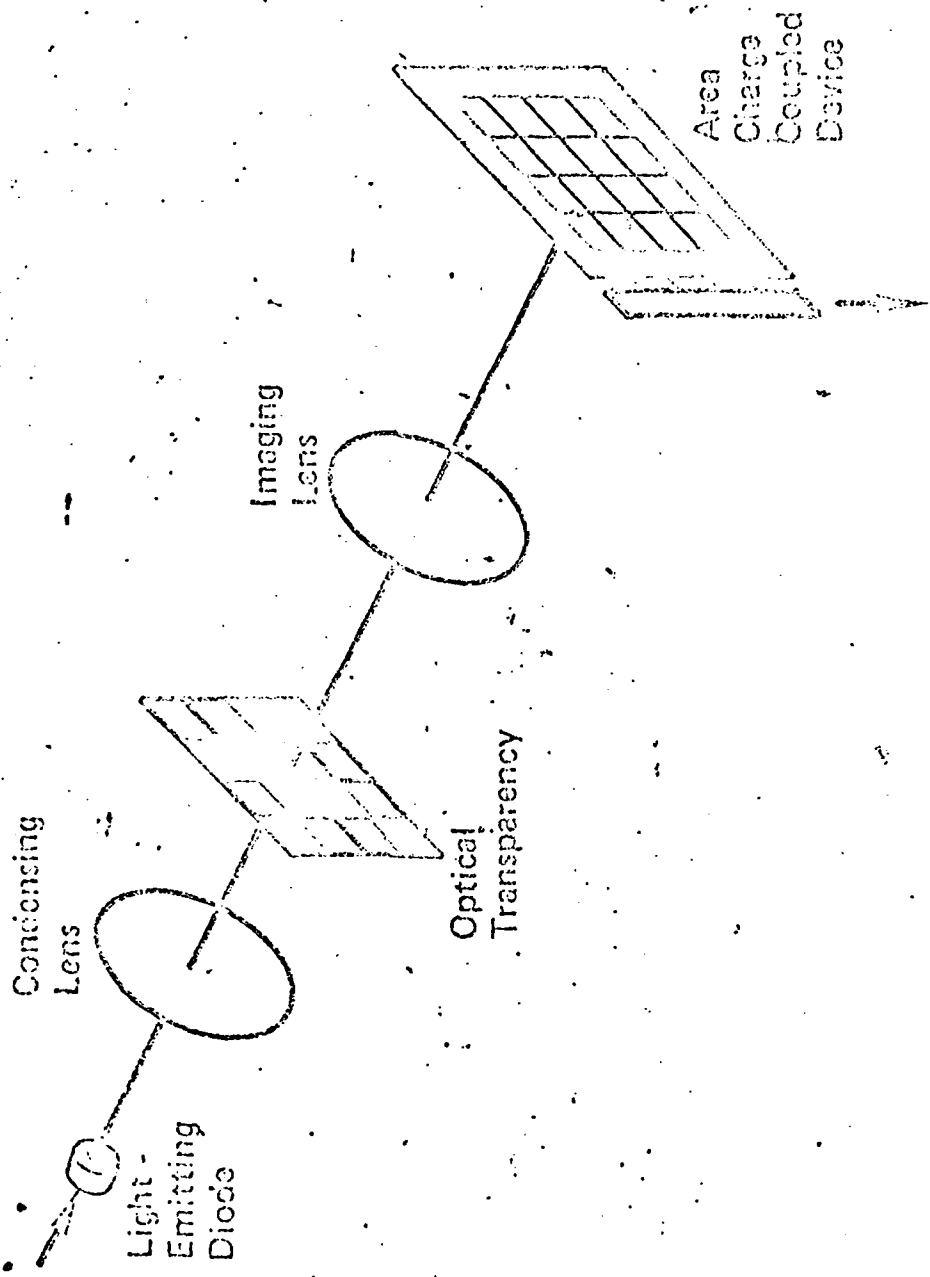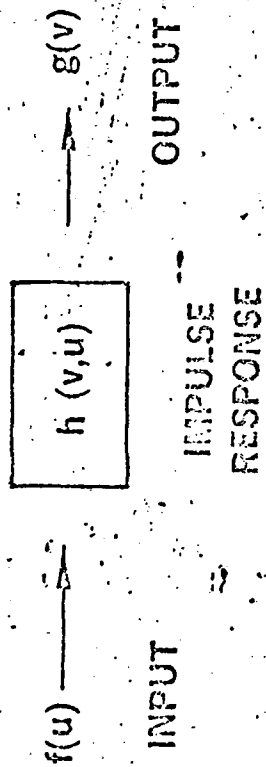
CONTINUOUS $g(v) = \int h(v,u)\,f(u)\,du$

DISCRETE $g_m = \sum_{n=1}^{N} h_{mn}\,f_n$, $m = 1,2,\ldots M$.

MATRIX

$$\begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_M \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1N} \\ h_{21} & h_{22} & & \\ \vdots & & & \vdots \\ h_{M1} & & \cdots & h_{MN} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix}$$

Attachment 12

PROGRAMMABLE ELECTRO-OPTICAL PROCESSOR
EXAMPLES OF IMPULSE RESPONSES

CONVOLUTION
$$h(v,u) = h(v-u)$$

CROSS-CORRELATION
$$h(v,u) = h(u-v)$$

AUTO-CORRELATION
$$h(v,u) = f(u-v)$$

FOURIER TRANSFORM
$$h(v,u) = e^{-i2\pi uv}$$

COSINE TRANSFORM
$$h(v,u) = \cos(2\pi uv)$$

SINE TRANSFORM
$$h(v,u) = \sin(2\pi uv)$$

LAPLACE TRANSFORM
$$h(v,u) = e^{-vu}$$

HANKEL TRANSFORM
$$h(v,u) = 2\pi J_0(2\pi vu)\,u$$

MELLIN TRANSFORM
$$h(v,u) = u^{v-1}$$

ABEL TRANSFORM
$$h(v,u) = 2u/(u^2 - v^2)^{1/2}$$

HILBERT TRANSFORM
$$h(v,u) = \frac{1}{\pi}\,\frac{1}{(u-v)}$$

HARTLEY TRANSFORM
$$h(v,u) = \cos(2\pi uv) + \sin(2\pi uv)$$

Attachment 12

Attachment 12

6
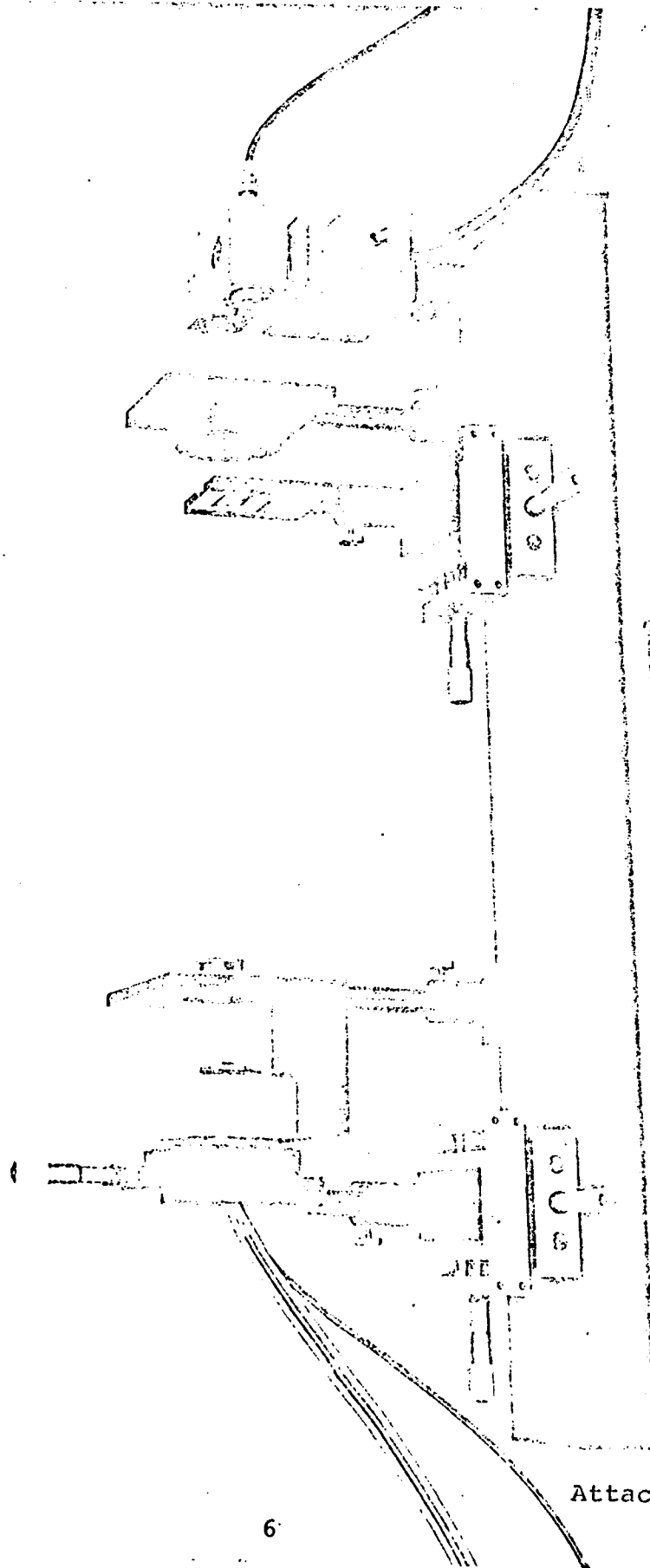
WALSH TRANSFORM MASK
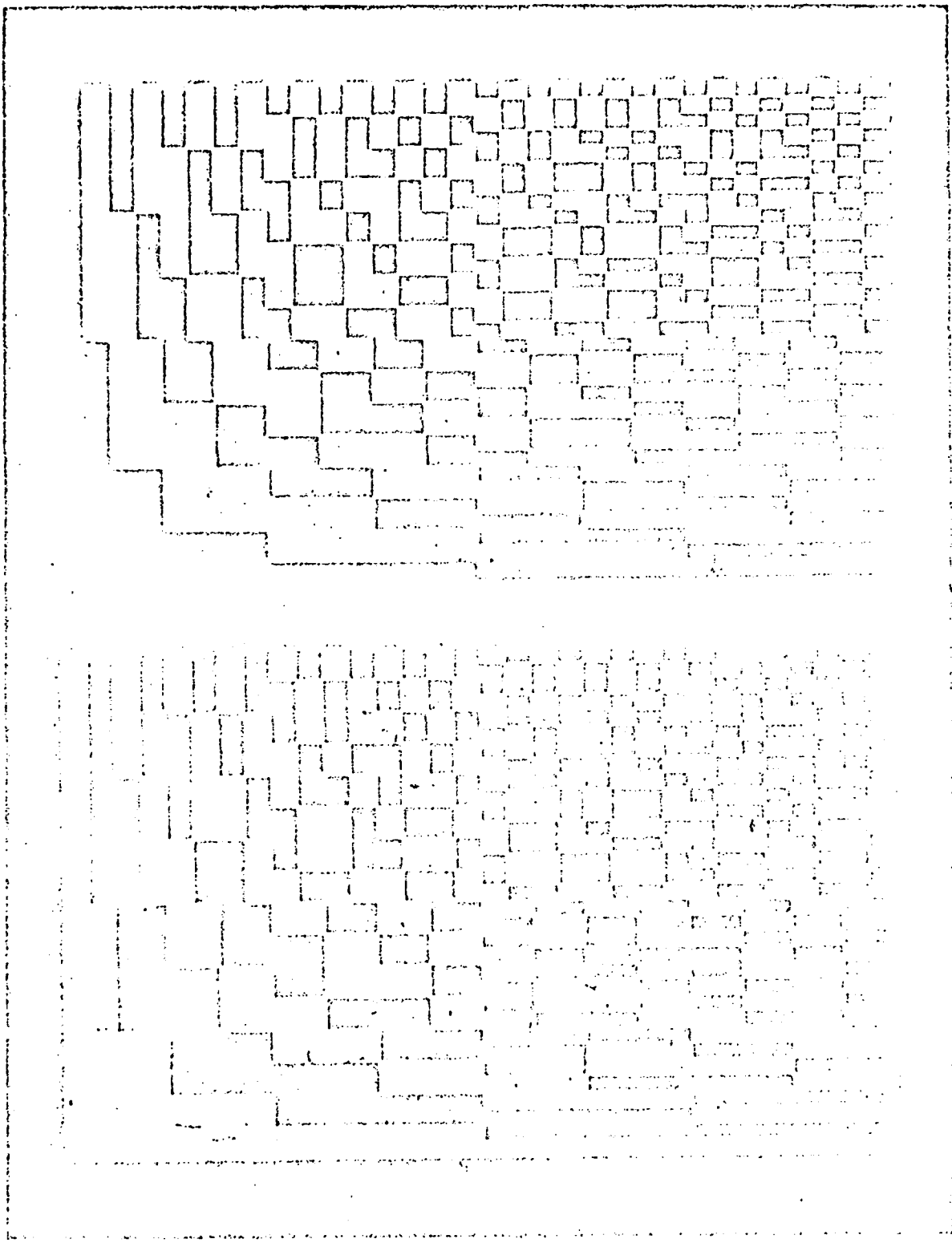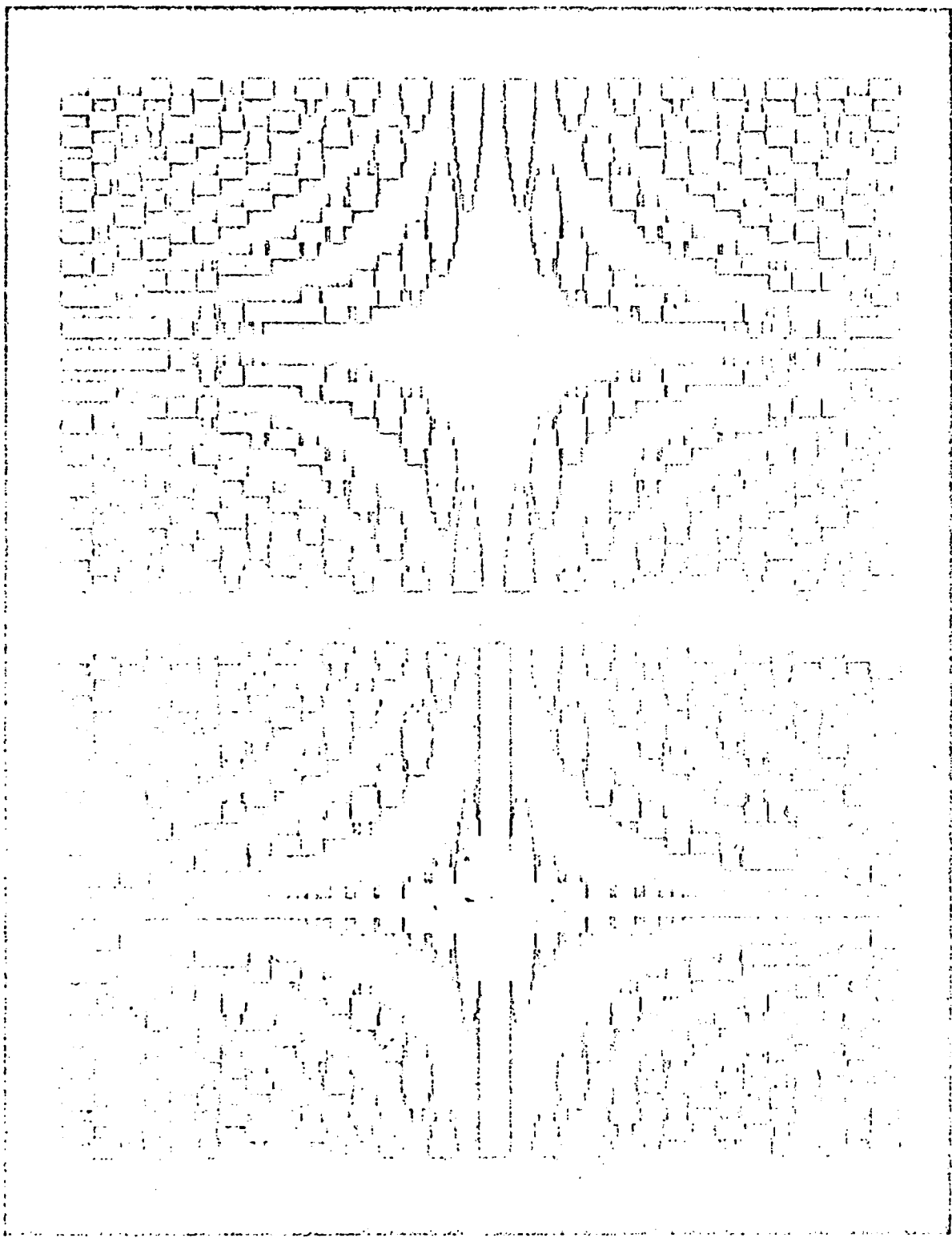
COSINE TRANSFORM INPUT-OUTPUTS

Input-A

Input-B

Positive
Output

Negative
Output

Positive
Output

Negative
Output

Cosine Transform

Cosine Transform

Input-C

Input-D

Positive
Output

Negative
Output

Positive
Output

Negative
Output

Cosine Transform

Cosine Transform
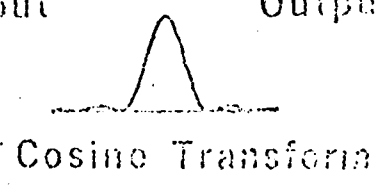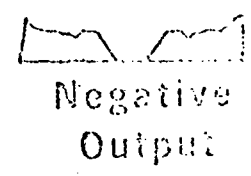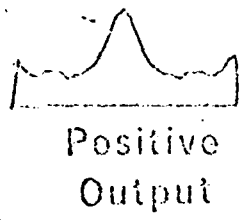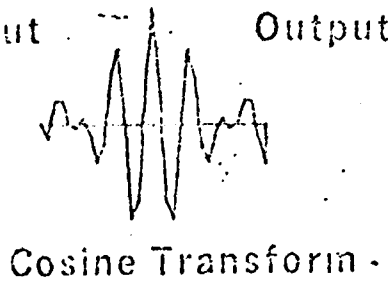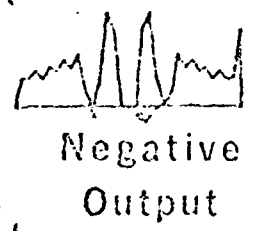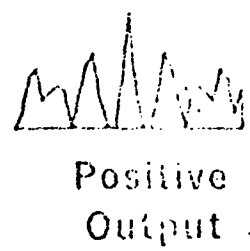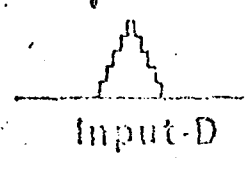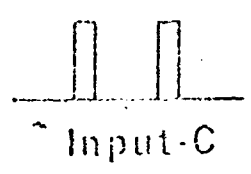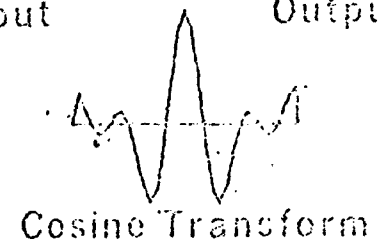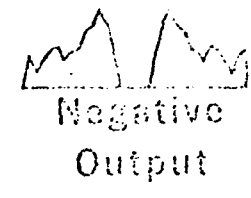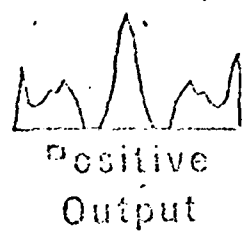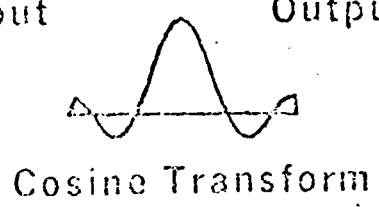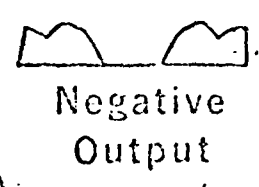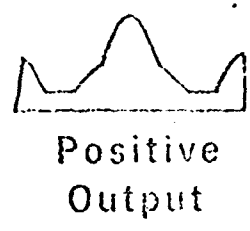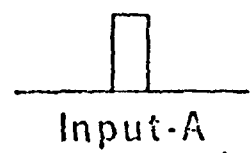
NELC

An optical incoherent correlator

KEITH BROMLEY

Electro-Optics Technology Division,
Naval Electronics Laboratory Center, San Diego, California 92152

Abstract. This paper will discuss a technique for cross-correlating a one-dimensional input signal with a library of stored reference signals simultaneously. The technique involves the linear scanning of a temporally-modulated image of a photographic reference mask across a temporally-integrating read-out device. First, a physical feeling for the technique is given, followed by a mathematical analysis, a description of a breadboard system now in operation, a simple experiment to demonstrate the correlators features, and a mention of some alternate configurations desirable for specific applications.

1. Introduction

This paper reports the development of an optical correlator capable of the real-time cross-correlation of an incoming electronic one-dimensional signal with each member of a stored reference library of one-dimensional signals. A novel feature of the device is that it avoids the real-time input transducer problem, which has plagued coherent optical processing systems, by introducing the input signal in the form of a temporal intensity modulation of an incoherent light source. While this device does not have the versatility of a coherent optical processor, the ability of this device to perform parallel cross-correlations of an incoming one-dimensional signal with a stored reference library in real-time and in an inexpensive, compact system makes it very attractive for many applications. Several alternate designs for incoherent optical correlators have previously been reported in the literature [1, 2, 3, 4].

2. Principle of operation

Figure 1 illustrates the basic concept. First, a light source S is imaged by a condensing lens C into the entrance aperture of an imaging lens L. In this light beam, immediately after the condensing lens, is placed a photographic transparency which contains the library of reference signals of interest. This mask has the form of a linear array of N horizontal channels with each channel having a different spatial variation in intensity transmittance corresponding to some reference signal. Lens L images this transparency into an output plane P. Between the mask and the imaging lens L is placed a rocking mirror M which causes the image to repetitively translate with constant velocity v across the plane P.

Assuming that the intensity transmittance along the ith channel of the mask exactly matches the input signal, then, for some velocity v of the output image, there will exist a point in the image of the ith channel in the output plane at which the temporal intensity variation due to the moving image is coincident with the input signal modulating the light source. That is, transparent areas of

c2
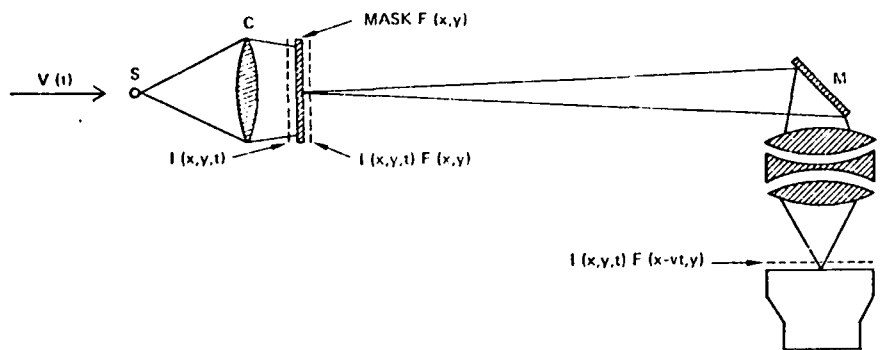
Figure 1. Basic arrangement of optical incoherent correlator.

that channel on the mask will be coincident with high light levels from the modulated light source and opaque areas will be coincident with low light values. At other points along that channel or at any point along different channels, some opaque areas will be associated with high light values, and vice versa. If the intensity pattern in the output plane is now integrated over one complete scan of the mask (e.g. by means of a vidicon tube, photographic film, or the eye), then a correlation peak appears at that one point along the *i*th channel.

3. Mathematical analysis

In formal terms, the operation of the correlator in detecting and classifying a signal in a noise background can be described as follows. Let the incoming signal to be identified be denoted by $S(t)$. Assume either that the signal exists for only some finite time interval T or, if it is continuous, that we are only considering a segment of it which is T in length. If the signal is bipolar, then in order to represent it as a time-varying intensity field, it must be added to a d.c. bias since negative light intensity is not d-fined. Thus the signal modulating the light source is

$$V(t) = B + KS(t/a) + N(t), \tag{1}$$

where $S(t)$ is assumed to have a normalized peak value of unity, K and a allow for possible scale changes in amplitude and frequency respectively, and $N(t)$ is the noise background.

Thus the light field intensity incident on the reference mask is

$$I(x, y, t) = I(x, y)V(t) = I V(t), \tag{2}$$

where it is assumed that the filter illumination function has the space-independent value I.

The light emerging from the *i*th channel of the mask is

$$IV(t)F_i(x), \tag{3}$$

where $F_i(x)$, the intensity transmittance of the *i*th channel, is also a biased function in order for bipolar signals to be recordable as an intensity transmittance on a medium such as photographic film. So $F_i(x)$ is written as

$$F_i(x) = B_i + K_i R_i(x/b_i), \tag{4}$$

where b_i is a measure of the scale of the recorded version of the reference signal $R_i(x)$, K_i is an amplitude scale factor, and again $R_i(x)$ is assumed normalized to a peak value of unity.

Attachment 12

Assuming perfect unit magnification imagery [5] via the rocking mirror, the light field intensity incident on the vidicon face from the ith channel at time t in a single scan is

$$0_i(x, t) = IV(t)F_i(x - vt - \Delta), \qquad (5)$$

where v is the constant velocity of the mask image as it scans across the vidicon face, and Δ is the phase of the scan which is controlled by the mirror-drive electronics. At the output plane, a vidicon or some other means is used to integrate $0_i(x, t)$ over a single scan time T which corresponds to the length of time required for the mirror to scan across the output plane a total distance equal to the x-dimension of the mask image.

The integrated output image of the ith channel is therefore

$$0_i(x) = \int_{-T/2}^{T/2} 0_i(x, t)\, dt$$

$$= I \int_{-T/2}^{T/2} V(t) F_i(x - vt - \Delta)\, dt. \qquad (6)$$

Because of the required biasing, writing this integral in terms of the input signal $S(t)$ and the reference function $R(x)$ results in six separate terms:

$$0_i(x) = IBB_iT$$

$$+ IBK_i \int_{-T/2}^{T/2} R_i\left(\frac{x - vt - \Delta}{b_i}\right) dt$$

$$+ IB_iK \int_{-T/2}^{T/2} S(t/a)\, dt$$

$$+ IKK_i \int_{-T/2}^{T/2} S(t/a)R_i\left(\frac{x - vt - \Delta}{b_i}\right) dt$$

$$+ IB_i \int_{-T/2}^{T/2} N(t)\, dt$$

$$+ IK_i \int_{-T/2}^{T/2} N(t)R_i\left(\frac{x - vt - \Delta}{b_i}\right) dt. \qquad (7)$$

The fourth term is the one of interest. On rearranging the argument of the second factor of the integrand this term becomes

$$IKK_i \int_{-T/2}^{T/2} S(t/a)R_i\left[\frac{t - [(x - \Delta)/v]}{(-b_i/v)}\right] dt \qquad (8)$$

and resembles a correlation integral.

Now consider the case where the ith channel is matched to the incoming signal, i.e. $R_i(x) = S(x)$. Then equation (8) describes the correlation output term for the matched condition in the ith channel.

A reference scale search can be accomplished by repeatedly modulating the light source with this incoming signal while varying the angular velocity of the mirror for each successive scan. Matching will be achieved when

$$v = -\frac{b_i}{a}. \qquad (9)$$

After a simple change of variables, and noting that $S(t)$ is a real function, expression (8) becomes proportional to the autocorrelation function.

Attachment 12

$$\phi_{ii}\left(\frac{x-\Delta}{av}\right)=aIKK_i\int_{-T/2a}^{T/2a}S(t')S\left[t'-\left(\frac{-\Delta}{v}\right)\right]dt'. \qquad (10)$$

For any other (i.e. non-matched) channel of the mask, the output term becomes proportional to the cross-correlation function between the input and reference signals.

$$\phi_{ij}\left(\frac{x-\Delta}{av}\right)=aIKK_j\int_{-T/2a}^{T/2a}S(t')R_j\left[t'-\left(\frac{x-\Delta}{av}\right)\right]dt'. \qquad (11)$$

In addition to the desired correlation term in equation (7) there are five additional terms which in general will reduce the desired output signal of the vidicon. The first term is a constant which can be removed electrically at the vidicon itself.

The second and third terms of equation (7) involve integrations over the reference and signal functions respectively. Although the third term is a constant for a given signal, the second term is a function of the quantity $(x-\Delta)/v$. Thus a given value of $x=\Delta$ determines the amount of the reference function R_i which is actually integrated (for the case of a finite signal of length T). However, the peak value of this degrading term can be found by performing the integration for $x-\Delta$. Parks [1] has stated that in the case of a 511-bit maximal length binary-coded signal the second and third terms are at least 30·6 dB below the correlation peak. (This figure will, of course, vary according to the signal type.) Little can be said of the fifth and sixth terms without specifying some statistics of the noise. For many applications, the noise has zero mean and when integrated over a sufficiently large interval these terms can be considered negligible.

4. Experimental arrangement

A laboratory model of this correlator is shown in the photograph of figure 2. The optical components are mounted in Kinematic Electro-Optical Construction modules to facilitate alignment. The light source is a Monsanto MV4 light-emitting diode with a peak output of about 1 mW at 6700 Å. The photographic film used as the 1 in. square reference mask is Eastman Kodak 649F developed to manufacturer's recommendations.

The 0·5 in. diameter mirror is mounted on a General Scanning Inc. galvano-meter model G-108. The imaging lens is a Schneider Componon 50 mm f/4·0, positioned to image the mask on to the face of a standard 525-line closed-circuit television system with a 4 : 1 minification. Driving the mirror-galvanometer system with a saw-tooth waveform results in a repetitive linear translation of the mask image across the vidicon face.

With the same input signal repetitively scanned across the vidicon at rates of 10 scans/sec or faster, the television monitor and the eye combine to perform the integration adequately. At slower rates, other integrators such as photographic film or a storage vidicon are needed.

5. Experiment description

A photographic transparency shown in figure 3 (*a*), containing twenty-five 90-bit pseudo-random binary codes, was inserted into the mask plane. The input signal modulating the light source was the output of a waveform synthe-sizer repetitively generating the code identical to the code contained along the
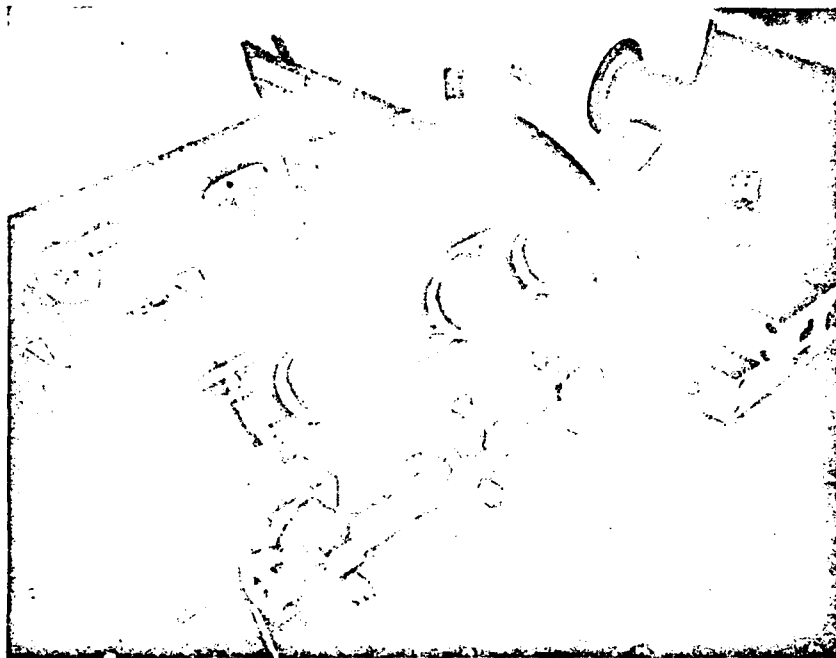
Figure 2. Photograph of experimental correlating system.

fifth channel of the mask. No noise was inserted into the system. Binary signals were used strictly for their convenience in generation and for their easily calculable correlation functions; the system works equally well with analogue waveforms.

Shown in figure 3 (*b*), at a scanning rate of 100 scans/sec, is the integrated output-plane image, recorded by simply exposing photographic film to the output-plane image for several scans. The bright autocorrelation peak in the fifth channel verifies the match.

By placing the television system in the output plane and observing, on an oscilloscope, the output waveform corresponding to the fifth channel, one sees the autocorrelation function shown in figure 3 (*c*). One can show that, within experimental error—primarily limited by the television system bandwidth and non-linearity of the mirror motion—this is exactly the autocorrelation function of the specified 90-bit code of 0's and 1's.

6. Alternate configurations

With the system just described, the vertical position of the correlation peak identifies the channel giving a match, the horizontal position indicates the phase Δ between the saw-tooth waveform driving the mirror and the incoming signal, and the relative intensity of the peak provides a measure of confidence in the identification of the match.

In many applications (e.g. where Δ is known *a priori* to be a fixed constant) the phase information is of no interest, thus leaving the horizontal position free

Attachment 12

K. Bromley

(a)

(b)
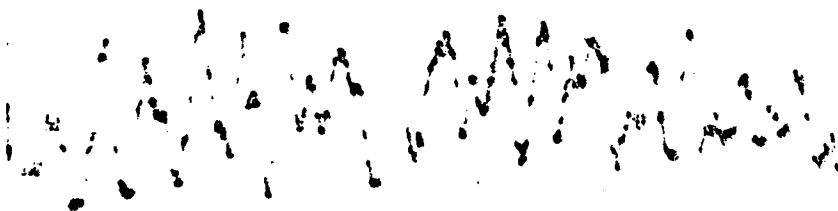
(c)

Figure 3. (a) Reference mask containing twenty-five 90-bit codes. (b) Integrated output-plane image showing autocorrelation peak in centre of fifth channel. (c) Oscilloscope trace of video signal corresponding to fifth channel.

for other uses.   In some of these applications it is desirable to utilize this relaxation of a constraint to achieve freedom from having to perform a scale search. One way to do this is to input both the input signal and the reference signal in log space (to any base) rather than linear space.   In other words, the input to the light source is a functiin of log time rather than time, and the reference channels along the mask are functions of log space rather than space [6].   Since

$$\log \alpha t = \log \alpha + \log t \tag{12}$$

any scale shift in time (such as a Doppler shift in the input signal) manifests itself as a simple translation of the input signal in log space; and hence the horizontal position of the correlation peak will yield the Doppler shift.   No scale search is necessary.

In other applications where the input waveform is known *a priori*, the vertical position is left free to be used for a scale shift.   This is done by simply letting every channel of the mask be the same waveform, but different scale shifts thereof.   The vertical shift of the peak thus represents scale shift, and again the horizontal distance represents phase.

Cet article discute une technique pour la corrélation simultanée d'un signal d'entrée unidimensionnel avec une bibliothèque de signaux de référence emmagasinés.   Cette technique fait appel au balayage linéaire de l'image modulée temporellement d'un masque photographique de référence à travers un système de lecture intégrant temporellement. On présente d'abord les bases physiques de cette technique et ensuite une analyse mathématique, une description du système qui fonctionne actuellement, une expérience simple pour démontrer les caractéristiques des corrélateurs, et une mention de quelques autres configurations utiles pour des applications spécifiques.

Diese Arbeit diskutiert eine Technik zur simultanen Kreuzkorrelation eines eindimensionalen Eingangssignals mit einer Bibliothek von gespeicherten Referenzsignalen.   Die Technik beinhaltet die lineare Abtastung des zeitmodulierten Bildes einer photographischen Referenzmaske über eine zeitintegrierende Auslesevorrichtung. Zuerst wird ein physikalisches Gefühl für die Technik vermittelt, gefolgt von einer mathematischen Darstellung sowie der Beschreibung einer jetzt arbeitenden Versuchsausführung und einem einfachen Experiment zur Demonstration der besonderen Merkmale des Korrelators. Schließlich werden einige alternative Anordnungen erwähnt, die für bestimmte Anwendungen wünschenswert sind.

REFERENCES

[1] PARKS, J. K., 1965, *J. acoust. Soc. Am.*, **37**, 368.
[2] TALAMINI, A. J., Jr., and FARNETT, E. C., 1965, *Electronics*, **38**, 58.
[3] JACKSON, D. E., 1966, *Sperry Engineering Review*, **19**, 15.
[4] CHANG, M., and McCRICKARD, J. T., 1971, *Appl. Optics*, **10**, 2784.
[5] Unit magnification, although not necessary, is assumed here for simplicity.
[6] Actually, three parameters can be utilized for this purpose; the rate at which the input signal modulates the light source, the spatial representation of the reference signal along the mask channel, and the image speed across the vidicon.   If any two of them are made logarithmic in time (or space, in the case of the mask) and the other made linear then this technique works.

Attachment 12

Matrix Multiplication Using Incoherent Optical Techniques

Richard P. Bocker

The use of incoherent electrooptical analog methods for performing matrix-vector multiplication has been investigated mathematically. A technique for encoding the matrix information on a two-dimensional binary optical transparency by means of an area modulation scheme is described. The one-dimensional discrete finite Fourier transform, viewed from the standpoint of matrix-vector multiplication, has been performed experimentally to demonstrate feasibility. Matrix and vector array sizes employed were 33 × 33 and 33 × 1, respectively. The average value of the correlation coefficients between theoretically derived and experimental data was found to be 0.95.

Introduction

The performance of matrix multiplication with coherent optical analog methods has previously been reported in the literature.[1-3] Presented in this paper is a description of a technique utilizing incoherent technology for performing matrix-vector multiplication of the form

$$c_m = \sum_{n=1}^{N} a_{mn} b_n, \quad m = 1, 2, \dots, M. \quad (1)$$

The elements a_{mn} constitute an $M \times N$ matrix (A), whereas the elements b_n and c_m represent an $N \times 1$ column vector (B) and an $M \times 1$ column vector (C), respectively. The realization of the matrix-vector multiply operation defined by Eq. (1) is possible with the use of an incoherent optical correlating device previously developed.[4] Presently, this incoherent optical device is limited to doing matrix-vector multiply operations when the column vector (B) is real-positive (the matrix (A) may be complex). Although the device is not as versatile as a coherent optical processor, it is nevertheless well suited for many signal-processing applications involving various discrete linear transformations.

System Description

Figure 1 depicts the incoherent electrooptical system used to perform the matrix-vector multiply operation. The system consists of a modulatable light source (s), a condensing lens (l_1), an optical transparency (m), a scanning mirror (r), an imaging lens (l_2), and an integrating detector (d). A time sequence of electrical pulses, containing the vector (B)

The author is with the U.S. Naval Electronics Laboratory Center, San Diego, California 92152.

information, intensity-modulates the light source. The condensing lens maximizes the light throughput in the system by imaging the light source into the entrance pupil of the imaging lens. Directly behind the condensing lens is placed the optical transparency that contains the matrix (A) information. An image of the optical transparency is formed at the detector face by the imaging lens. Between the optical transparency and the imaging lens is placed a scanning mirror that causes the image of the transparency to repetitively translate with a constant velocity across the detector face. The output column vector (C) information is generated at the integrating detector.

The light source is a Monsanto MV4 light-emitting diode with a peak output of about 1 mW centered at 670 nm. The optical transparency is a 35-mm slide made from Kodak high-contrast Kodalith film. A 1.25-cm-diameter mirror is mounted on a General Scanning Inc. galvanometer, model G-108. The imaging lens is a Schneider Componon 50-mm $f/4$ that is positioned to image the transparency onto the face of a standard 525-line closed-circuit television vidicon with a minification of 4:1 between object transparency and image. Driving the mirror-galvanometer system with a sawtooth electrical waveform results in a repetitive linear translation of the image across the vidicon face.

Mathematical Preliminaries

The fundamental equation of interest is the imaging equation that connects the exposure at the detector plane in terms of the irradiance of the light field incident on the transparency and the intensity transmittance of the optical transparency. If the effects of lens aberrations and diffraction are neglected, the exposure is given by the following superposition integral:
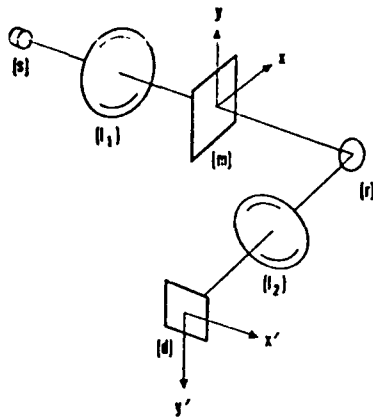
Fig. 1. Layout diagram of the incoherent electrooptical matrix-vector multiplier. Following are system parts: *(s)* incoherent modulatable light source; *(l₁)* condensing lens; *(m)* optical transparency containing the matrix (A) information; *(r)* scanning mirror; *(l₂)* imaging lens; *(d)* integrating detector.

$$E(x',y') = (1/m_t{}^2) \int_{-\infty}^{\infty} I(t)\, \tau(x'/m_t,\, y'/m_t + vt/m_t)\, dt. \quad (2)$$

$I(t)$ is the irradiance of the incident light field and $\tau(x,y)$ is the intensity transmittance of the optical transparency. The quantity m_t represents the transverse magnification between the transparency and detector planes. The velocity at which the image of the transparency is scanned across the detector face is given by v. It is the function $I(t)$ that contains the input vector (B) information and the function $\tau(x,y)$ that contains the matrix (A) information. As will be shown, the output vector (C) information is contained in the function $E(x',y')$.

The vector (B) information is input as an electrical time sequence of rectangular pulses. This signal intensity-modulates the incoherent light source. Ideally, the light source and the condensing lens are configured in a manner such that the light beam incident on the transparency has a uniform irradiance distribution over the beam diameter, which varies in time according to the equation

$$I(t) = \sum_{k=1}^{N} b_k\, \text{rect}[(t - k\Delta t - t_0)/T]. \quad (3)$$

N represents the total number of light pulses, b_k the height of the kth pulse, Δt the spacing between adjacent pulse centers, T the pulse duration, and t_0 an arbitrary time shift. The rectangle functions appearing in Eq. (3) are defined by[5]

$$\text{rect}(t) \equiv \begin{cases} 1 & |t| < \tfrac{1}{2}, \\ 0 & |t| > \tfrac{1}{2}. \end{cases}$$

Figure 2 depicts the signal $I(t)$. As is evident from Eq. (3), the rectangular pulse heights contain the column vector (B) information. The present optical system configuration and encoding scheme constrain the vector (B) elements to take on only real-positive values.

The matrix (A) information is encoded on the op-

tical transparency in binary form. For simplicity, the discussion is limited to the case in which the elements of matrix (A) are real-positive. The more general case of a complex matrix (A) is treated in Appendix A. The intensity transmittance of the transparency is specified according to the equation

$$\tau(x,y) = \sum_{n=1}^{N} \sum_{m=1}^{M} \text{rect}[(x - m\Delta x - x_0)/W]$$
$$\times\, \text{rect}[(y - n\Delta y - y_0)/a_{mn}]. \quad (4)$$

The transparency contains a total of MN clear rectangular apertures arranged in a rectangular array as depicted in Fig. 3. There is one-to-one correspondence between each rectangular aperture in the array and each element in the matrix (A). The linear dimensions of the $(m\text{th},n\text{th})$ aperture in the array are given by W and a_{mn} in the x and y directions, respectively. W is the same for all apertures, whereas a_{mn} is equal to the $(m\text{th},n\text{th})$ element of matrix (A). The quantities Δx and Δy correspond to the spacing between aperture centers, and x_0 and y_0 represent arbitrary spatial shifts. The use of binary optical transparencies avoids many of the problems encountered in fabricating continuous tone analog masks. Binary masks have previously been employed in both holographic and coherent optical data processing applications.[6-8]

If the mathematical expressions for $I(t)$ and $\tau(x,y)$ are substituted into Eq. (2), it can be easily shown that the exposure can be written in the form

$$E(x',y') = \sum_{m=1}^{M} c_m(y')\, \text{rect}[(x'/m_t - m\Delta x - x_0)/W]. \quad (5)$$

The quantities $c_m(y')$ are defined by

$$c_m(y') \equiv \sum_{k=1}^{N} b_k\, \left\{ (1/vm_t{}^2) \sum_{n=1}^{N} \int_{-\infty}^{\infty} \text{rect}[(y'' - kv\Delta t - vt_0)/vT] \right.$$
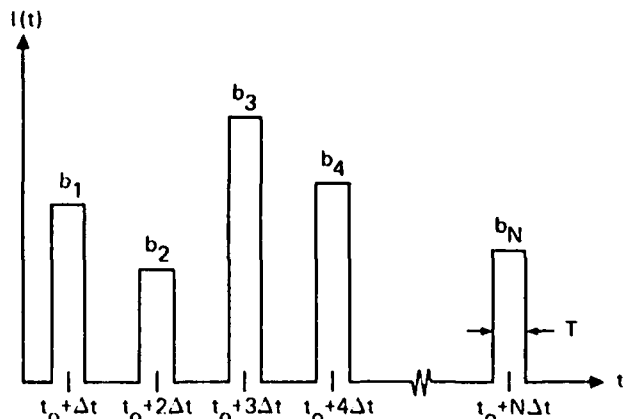$$\left. \times\, \text{rect}[(y''/m_t + y'/m_t - n\Delta y - y_0)/a_{mn}]\, dy'' \right\}.$$

Fig. 2. Typical temporal light signal $I(t)$ containing input column vector (B) information, used to illuminate the optical transparency.
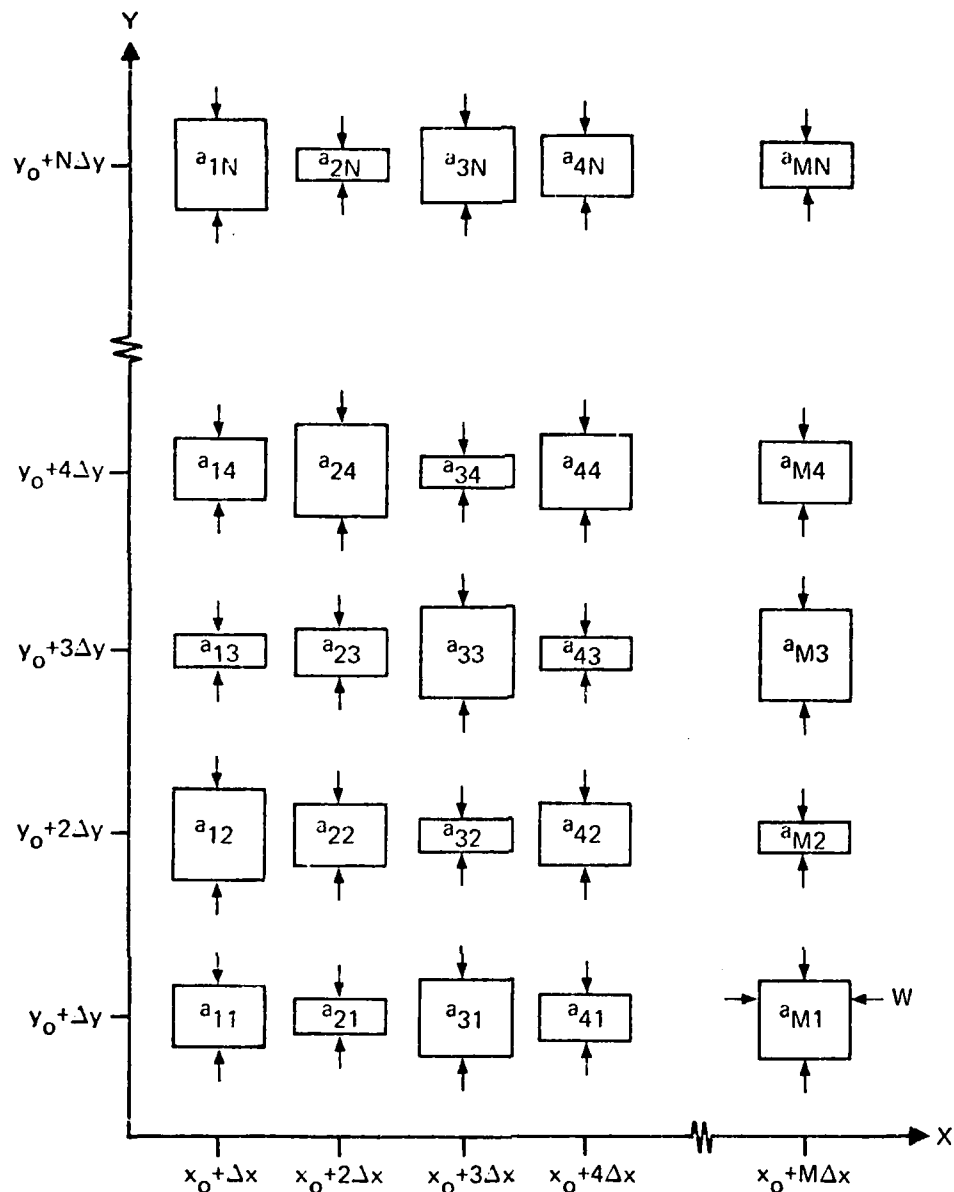
Attachment 12

Fig. 3. Area modulation scheme employed for encoding the matrix (A) information on the optical transparency.

where the new variable of integration, y'', is equal to vt. In Appendix B it is shown that the quantities $c_m(y')$, when evaluated at $y' = 0$, yield a set of coefficients that are linearly proportional to the elements of the output vector (C) defined in Eq. (1). The result is

$$c_m(0) = (1/vm_t)\sum_{n=1}^{N}a_{mn}b_n. \qquad (6)$$

Aside from the constant factor $(1/vm_t)$, Eq. (6) is identical to Eq. (1). Referring to Eq. (5), we see that the exposure at $y' = 0$ contains the output column vector (C) information in terms of a spatial sequence of rectangular pulses. M represents the total number of pulses, $c_m(0)$ the height of the mth pulse, $\Delta x m_t$ the spacing between pulse centers, Wm_t the spatial width of each pulse, and $x_0 m_t$ an arbitrary spatial shift. Figure 4 graphically depicts the exposure for $y' = 0$.

Experimental Results

A set of experiments was performed with the incoherent optical system previously described to demonstrate the matrix-vector multiply operation. The particular matrix-vector multiply operation demonstrated was that of a discrete finite Fourier transform. The matrix (A) associated with this transformation is a square $N \times N$ matrix whose elements are given by

Attachment 12

Fig. 4. Exposure function $E(x',y')$ at $y' = 0$ containing the output column vector (C) information.

Fig. 5. Photograph of the 35-mm optical transparency used to compute a 33-point discrete Fourier transform.

$$a_{mn} = \exp\{-j2\pi[m-(N+1)/2][n-(N+1)/2]/(N-1)\}. \tag{7}$$

The integer N may take on only odd values. For this set of experiments it was equal to 33. The input column vector (B) contains sampled values of the temporal function to be Fourier-transformed. The elements of the output column vector (C) are the Fourier transform coefficients. Shown in Fig. 5 is a photograph of the optical binary transparency used to perform the discrete finite Fourier transform operation for $N = 33$. As discussed in Appendix A, the transparency actually contains four masks, one associated with each of the four real-positive matrices required to describe the complex matrix (A).

Three different experiments were performed with the finite Fourier transform mask. Shown in Fig. 6(a), (b), and (c) are theoretically predicted outputs for different inputs. The first line in each of these figures represents the input column vector (B) infor-

mation presented as a sequence of 33 rectangular pulses. The column vectors associated with each of the three different inputs are given by

$$(B) = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.2 \\ 0.4 \\ 0.6 \\ 0.8 \\ 1.0 \\ 0.8 \\ 0.6 \\ 0.4 \\ 0.2 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \text{ and } \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \\ 0.5 \\ 0.6 \\ 0.7 \\ 0.8 \\ 0.9 \\ 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}.$$

The four curves appearing in the second line of Fig. 6(a), (b), and (c) correspond to theoretically predicted curves of the exposure $E(x',y')$ evaluated at $y' = 0$. For completeness, the third line of these same figures corresponds to the real and imaginary curves associated with the real and imaginary parts of the column vector (C). The real (imaginary) curve in the third line is obtained by subtracting the real (imaginary) negative curve from the real (imaginary) positive curve in the second line of the same figure.

For these experiments, the input signals and optical transparency were designed so that the temporal pulse duration T was equal to the pulse spacing Δt, the aperture width W was equal to the horizontal aperture spacing Δx, and the maximum value of the moduli of the quantities a_{mn} was equal to the vertical aperture spacing Δy. Because W is equal to Δx, the spacing between adjacent spatial pulses making up the outputs is zero. This is the reason that the outputs appear to be continuous curves rather than discrete sets of pulses.

Figure 7(a), (b), and (c) are experimental curves associated with the theoretical curves appearing in the second line of Fig. 6(a), (b), and (c), respectively. These experimental curves correspond to actual cathode-ray-tube traces of that one line of vidicon associated with $y' = 0$ recorded on Polaroid film.

A set of correlation coefficients was computed to determine how well the experimental results agreed with theoretical predictions. The correlation coefficient K associated with two sets of data $\{x_m\}$ and $\{y_m\}$ is defined by[9]

Attachment 12

Fig. 6. Theoretically predicted output curves containing the column vector (C) information for various inputs.

Attachment 12

Fig. 7. Experimental results: curves *(A)*, *(B)*, and *(C)* are experimental outputs associated with theoretically predicted curves in the second line of Fig. 6(A), (B), and (C), respectively.

$$K = \sum_{m=1}^{M}(x_m-\bar{x})(y_m-\bar{y}) \Big/ \Big[\sum_{m=1}^{M}(x_m-\bar{x})^2\Big]^{1/2}\Big[\sum_{m=1}^{M}(y_m-\bar{y})^2\Big]^{1/2}.$$
(8)

where

$$\bar{x} = (1/M)\sum_{m=1}^{M} x_m$$

$$\bar{y} = (1/M)\sum_{m=1}^{M} y_m.$$

The quantities x_m, for example, would correspond to theoretical predicted values of the output column vector (C), whereas the quantities y_m would correspond to experimentally measured values of the output column vector (C). A correlation coefficient was determined for each of the outputs associated with the three different inputs. The average value of the set of correlation coefficients determined was found to be equal to 0.95.

Appendix A: Complex Matrix (A)

With only minor additions to previously described encoding techniques, the incoherent electrooptical device for performing matrix-vector multiplication between a real-positive matrix (A) and a real-positive vector (B) can be applied equally well to the case in which (A) is complex. In this section, the technique used for encoding the complex matrix (A) information on an optical transparency is presented.

Any arbitrary complex matrix (A) can be decomposed into a linear combination of four real-positive matrices. Formally, we can write

$$(A) = (A)_{rp} - (A)_{rn} + j(A)_{ip} - j(A)_{in},$$
(A1)

where j is equal to the square root of minus one. The matrices $(A)_{rp}$, $(A)_{rn}$, $(A)_{ip}$, and $(A)_{in}$ contain the real-positive, real-negative, imaginary-positive, and imaginary-negative information, respectively, about the complex matrix (A). The information associated with each of these real-positive matrices can be encoded by means of the area modulation scheme previously described onto one optical transparency containing four distinct masks arranged side by side in a linear array. Each of the four masks in the array is uniquely associated with one of the four real-positive matrices appearing in Eq. (A1). The result is a single optical transparency containing the complex matrix (A) information.

Inserting this transparency into the incoherent optical device will give rise to four real-positive serial outputs. The relationship between the four serial outputs, the input (B), and the four real-positive matrices associated with (A) is given by the following set of equations, in which the outputs are denoted by $(C)_{rp}$, $(C)_{rn}$, $(C)_{ip}$, and $(C)_{in}$:

$$\begin{aligned}
(C)_{rp} &= (A)_{rp}(B),\\
(C)_{rn} &= (A)_{rn}(B),\\
(C)_{ip} &= (A)_{ip}(B),\\
(C)_{in} &= (A)_{in}(B).
\end{aligned}$$
(A2)

The complex vector output (C) can be constructed from these real-positive outputs with the equation

$$(C) = (C)_{rp} - (C)_{rn} + j(C)_{ip} - j(C)_{in}.$$
(A3)

Appendix B: Evaluation of $c_m(y')$ for $y' = 0$

In this section it will be shown that the quantities $c_m(y')$ evaluated at $y' = 0$ yield a set of coefficients that are linearly proportional to the elements of the column vector (C) defined by Eq. (1). In order to obtain the results of interest, three constraints must be imposed. They are

$$v\Delta t = m_t \Delta y,$$

$$vt_0 = m_t y_0,$$

$$vT \geq m_t a_{mn} \quad \text{for all } m \text{ and } n.$$

Attachment 12

Recall that the signal $I(t)$ is composed of N rectangular pulses. This implies that the exposure function $E(x',y')$ is nothing more than a superposition of N spatially displaced images of the optical transparency. To physically realize the matrix-vector multiply operation requires that the relative displacement between adjacent images be equal to the quantity $m_t\Delta y$. This relative displacement is governed by both the velocity v of the moving image and the temporal spacing Δt between adjacent light pulse centers. The first constraint, therefore, ensures the correct registration between the N spatially displaced images.

The second constraint, which relates the temporal and spatial shift variables t_0 and y_0, guarantees the correct over-all vertical positioning of the N images relative to the Cartesian coordinate system x',y' located in the detector plane. Correct vertical positioning ensures that the quantities $c_m(y')$ will yield information about the output column vector (C) when evaluated at $y' = 0$.

The last constraint is intrinsically related to the scheme in which the matrix (A) and vector (B) information has been encoded. Each of the N pulses comprising the signal $I(t)$ yields an image of the optical transparency. Since each pulse has a finite time duration T, each of the N resulting images is degraded by linear image motion. Linear image motion degradation plays a key role in the realization of the matrix-vector multiply operation. To see this mathematically, we first evaluate the expression for $c_m(y')$ for $y' = 0$, subject to the first two constraints:

$$c_m(0) = \sum_{k=1}^{N} b_k \left\{ (1/vm_t^2) \sum_{n=1}^{N} \int_{-\infty}^{\tau} \mathrm{rect}[(y'' - km_t\Delta y - m_t y_0)/vT] \right.$$
$$\left. \times \mathrm{rect}[(y'' - nm_t\Delta y - m_t y_0)/m_t a_{mn}]dy'' \right\} . \quad (B1)$$

Each of the integrals appearing in Eq. (B1) can be evaluated in a straightforward manner. Formally, we obtain

$$\int_{-\infty}^{\infty} \mathrm{rect}[(y'' - km_t\Delta y - m_t y_0)/vT]$$
$$\times \mathrm{rect}[(y'' - nm_t\Delta y - m_t y_0)/m_t a_{mn}]dy''.$$

$$= \begin{cases} a_{mn}m_t\,\delta_{kn} & \text{for } vT \geq a_{mn}m_t \\ vT\ \ \delta_{kn} & \text{for } vT \leq a_{mn}m_t. \end{cases} \quad (B2)$$

δ_{kn} is the Kronecker delta function. The only case of physical interest occurs when $vT \geq a_{mn}m_t$ for all m and n, namely, the last constraint.

Imposing the last constraint reduces Eq. (B1) to the final desired result, namely,

$$c_m(0) = (1/vm_t) \sum_{n=1}^{N} a_{mn}b_n. \quad (B3)$$

References

1. R. A. Heinz, J. O. Artman, and S. H. Lee, Appl. Opt., **9**, 2161 (1970).
2. D. P. Jablonowski, R. A. Heinz, and J. O. Artman, Appl. Opt., **11**, 174 (1972).
3. L. J. Cutrona, Optical and Electrooptical Information Processing, J. T. Tippett et al., Eds. (M. I. T. Press, Cambridge, 1965), pp. 97–98.
4. K. Bromley, Opt. Acta, **21**, 35 (1974).
5. R. Bracewell, The Fourier Transform and Its Applications (McGraw-Hill, San Francisco, 1965).
6. B. R. Brown and A. W. Lohmann, Appl. Opt., **5**, 967 (1966).
7. A. W. Lohmann and D. P. Paris, Appl. Opt., **6**, 1739 (1967).
8. A. W. Lohmann and D. P. Paris, Appl. Opt., **7**, 651 (1968).
9. L. G. Parratt, Probability and Experimental Errors in Science (Wiley, New York, 1961).

NAVAL RESEARCH

REVIEWS

May-June 1974

SPECIAL ISSUE   *Naval Electronics Laboratory Center*

Attachment 12

Optical Data Processing
for Fleet Applications

R. P. Bocker, K. Bromley, and M. A. Monahan

Electro-optics Technology Division
Naval Electronics Laboratory Center

Because of the Navy's broad interest in signal processing, it is currently funding many programs to develop systems for performing transform and matrix operations. Potential areas of system application include Fourier spectral analysis of signals, vocoding and bandwidth compression of voice, and numerous applications to transversal filtering in radar and sonar signal processing. Most of the currently available systems for performing matrix transformations are all digital electronic systems that require either time-consuming sequential computations of each point in the matrix or large amounts of hardware to achieve some degree of parallelism in operation. Since most of these processors are "hard wired" to perform a particular sequential algorithm, they are capable of performing only one type of transformation. The electro-optical processor currently being developed by the Naval Electronics Laboratory Center is programmable to perform any one of a number of mathematical transformations at very high speed. The optical, fully-parallel nature of the device allows almost instantaneous computation of very large transformations in a unit that can be significantly smaller and less complex than existing systems.

The use of optics in signal processing introduces two important features. The first feature is an extremely fast multiplication rate—one analog value can be multiplied by another in the time required for light to pass through an optical transparency (about a picosecond). The second feature is the parallel processing capability—the two-dimensional nature of light propagation allows many one-dimensional operations to be performed simultaneously. Since the advent of the laser, much effort has been expended in applying *coherent* optical techniques to signal processing so that the separate control of amplitude and phase obtained could be utilized. However, in many applications, such efforts are being thwarted by (1) vibration sensitivity due to the interferometric nature of many techniques, and (2) the lack of a real-time input material sufficiently developed to a cost-effective, compact, off-the-shelf system. Fortunately, there are many worthwhile applications that do not require vibration insensitivity and real-time operation, and much promising research is being conducted to solve these problems in other applications.

Attachment 12

NELC researchers chose to pursue a different tack—to investigate *incoherent* optical techniques. Such techniques have the features of fast multiplication rate and parallel operation but do not have the problems of vibration sensitivity and the lack of real-time input capability. Results of the Center's program to date have been very promising (1-4). Two early exploratory development models of NELC's electro-optical processor were designed to cross-correlate simultaneously a "live" input signal with a large reference library of stored signals. These systems were applied to the problems of automatic passive sonar classification and active sonar detection and localization (1, 2).

Figure 1 shows the most recent implementation of the Center's electro-optical processor. This system is capable of a large variety of linear transformations and linear filtering operations. The basic concept of the technique is described as follows: The electrical input signal modulates the radiance of a light emitting diode (LED) as a
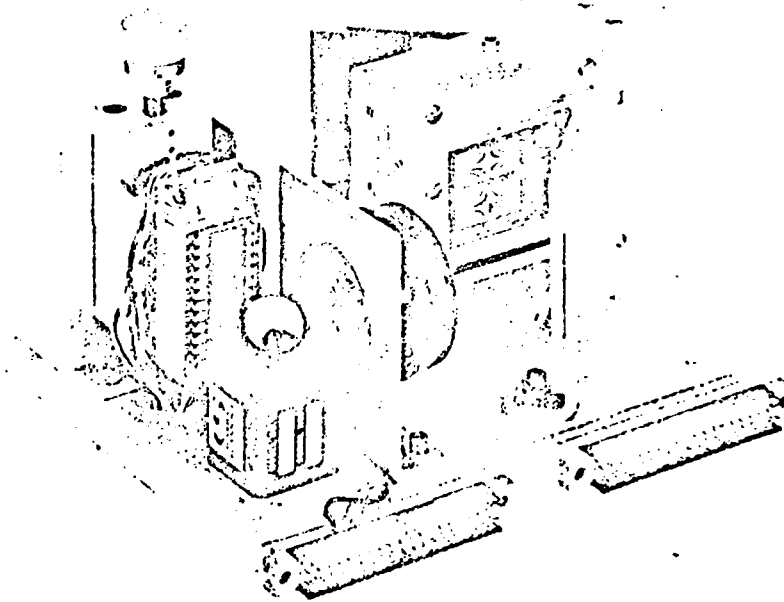
Figure 1 — Current implementation of NELC's programmable electro-optical processor. Components include the following: (1) light-emitting diode, (2) stationary mirror, (3) Fresnel condensing lens, (4) optical transparency, (5) stationery mirror, (6) imaging lens, (7) scanning mirror, and (8) linear charge-coupled device.

45

function of time. This LED is imaged by a Fresnel condensing lens into the entrance aperture of an imaging lens. In this light beam, immediately after the condensing lens, is placed a photographic transparency. This mask has the form of a linear array of horizontal channels with each channel having a different spatial vibration in intensity transmittance corresponding to some desired function. The imaging lens images this transparency, via a scanning mirror, onto a vertical row of integrating detectors—in the present embodiment, a linear charge-coupled device (CCD). The scanning mirror causes the image to repetitively translate horizontally, with constant velocity, across the face of the CCD array. The CCD integrates this intensity-modulated moving image during the mirror sweep. The resultant values are read out during the mirror's return.

In mathematical terms, the electrical input signal modulating the LED could be expressed as a column vector B of sampled data points, the mask as a matrix A, and the analog values serially read out of the CCD as a vector C. Then it can be shown that this device performs the vector-matrix multiply operation (4):

$$C = AB \quad \text{or} \quad c_m = \sum_{n=1}^{N} a_{mn} b_n, \quad m = 1, 2, 3, \ldots, M.$$

Some examples of operations that can be performed by this system are linear filtering, derivative operations, correlation, convolution, Fourier transforms, Laplace transforms, Walsh-Hadamard transforms, Z-transforms, and Mellin transforms. In fact, by simply replacing the photographic mask (i.e., the matrix A), it would be converted from, say, a Walsh transform device to a Z-transform device. Thus, it is broadly and simply programmable. Although space does not permit a description of how these masks are designed (4), an example is shown in Figure 2. This particular mask, reduced to 35 mm format, is the one used to generate the real and imaginary coefficients of a Fourier transform of the input data.

NELC is currently planning to build a processor incorporating a two-dimensional CCD array as the integrating and read-out device. With properly timed clocking sequences, the scanning operation can be performed entirely within the CCD chip, thereby eliminating the need for a scanning mirror. Such a system, composed of only an LED, a condensing lens, a replaceable mask, and a two-dimensional CCD array, will form an extremely compact, rugged, inexpensive system—with no moving parts—for performing vector-matrix

Figure 2 –
upper and l...
of the Four...
of each ma...
1 + sin ω...
different ω.

operations
limited to
off-the-she...
envisioned
be limited
with perist...

46
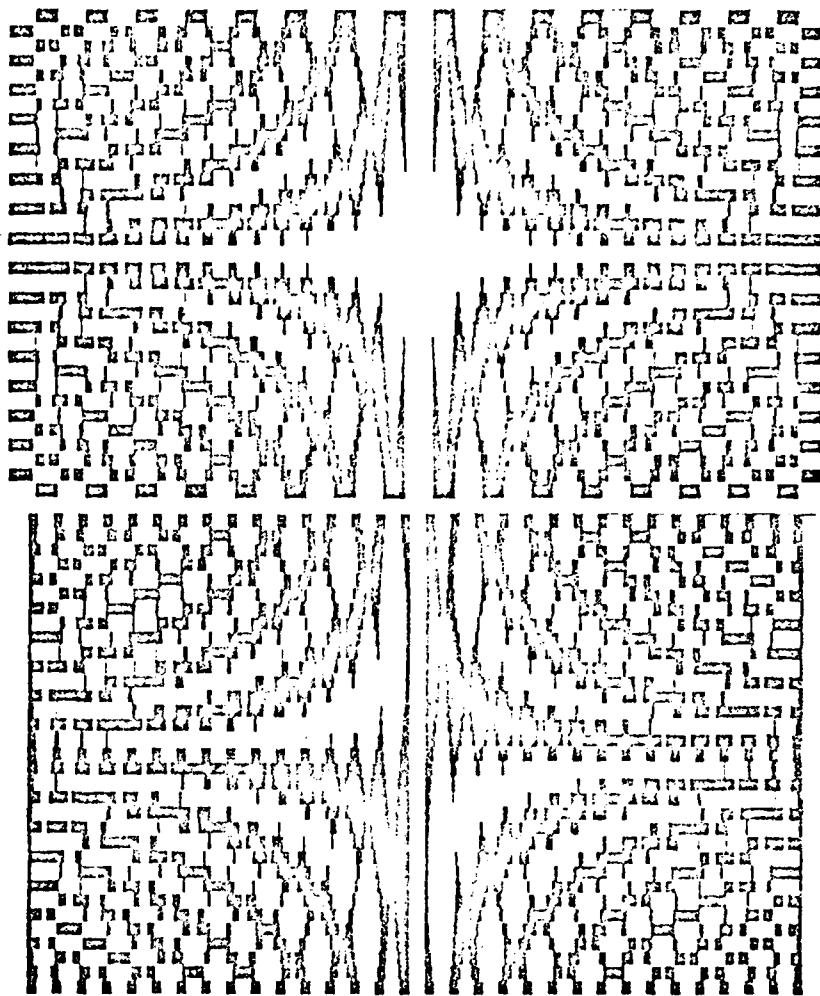
Figure 2 — When used in NELC's programmable electro-optical processor, the upper and lower halves of this mask produce the real and imaginary components of the Fourier transform coefficients of the input data. The horizontal length of each mask element is proportional to 1 + cos ω for the upper half, and 1 + sin ω for the lower half, with each horizontal line corresponding to a different ω.

operations at very high speed. The size of the matrix would be limited to that of the CCD array (100-by-100 element CCDs are off-the-shelf items today, and 1000-by-1000 element CCDs are envisioned within a few years). The fastest throughput rate would be limited to the readout rate of the CCDs (typically 10 MHz today, with peristaltic CCDs promising 1 GHz for the future (5)).

47

Attachment 12

References

1. Bromley, K., "An Optical Incoherent Correlator," *Optica Acta*, v. 21, no. 1, p. 35-41, January 1974

2. Naval Electronics Laboratory Center Report 1887, *Incoherent Optical Correlator for Active Sonar*, by T. C. Strand and C. E. Persons, 27 July 1973

3. Bocker, R. P., Bromley, K., and Monahan, M. A., *Incoherent Optical Data Processing*, paper presented at the Optical Society of America annual meeting, Rochester, N. Y., 11 October 1973

4. Bocker, R. P., "Matrix Multiplication Using Incoherent Optical Techniques," *Applied Optics*, to be published in August 1974 issue

5. Anonymous, "And Now—the PCCD," *Electro-optical Systems Design*, v. 6, no. 1, p. 6, January 1974

Improved Message Reproduction System

An improved message reproduction system was installed aboard USS LITTLE ROCK (CLG 4), the Sixth Fleet flagship, in July 1973, and is currently undergoing shipboard evaluation. Initial reports indicate that the system provides rapid, accurate reproduction and distribution (i.e., slotting) of shipboard message traffic and reduces manpower requirements. The off-the-shelf system was procured by NELC under a project sponsored by DNL's Navy Science Assistance Program (NSAP).

In late FY 72, COMSIXTHFLT requested NSAP assistance in solving problems experienced in the duplication and manual handling of messages aboard the flagship. A project team was established consisting of members from NELC (team leader); the Navy Publications and Printing Office, Washington; the Naval Ordnance Laboratory (NOL), White Oak; and the Naval Communications Station (NAVCOMMSTA), San Diego. The team's task was to analyze message-handling requirements aboard flag-configured ships, determine availability of equipment to satisfy those requirements, and make recommendations to the Navy for procuring equipments based on competitive evaluation.

In July 1972, team members conducted a survey to determine the density of message traffic associated with flagships of all sizes. Assistance was provided by the Atlantic and Pacific type commanders. In September 1972, NELC and NPPS personnel visited the CVAs on Yankee Station in the Tonkin Gulf to verify message loading at its heaviest peaks. From the data gathered, a specification for shipboard message reproduction and collating equipment was generated and three qualified systems were tested at NAVCOMMSTA San Diego in March 1973. The SIXTHFLT Staff then selected the configuration, from the qualified vendor's system, to be installed on LITTLE ROCK. A microfiche system for message storage offered by one bidder was also tested and accepted for shipboard use.

48

Attachment 12
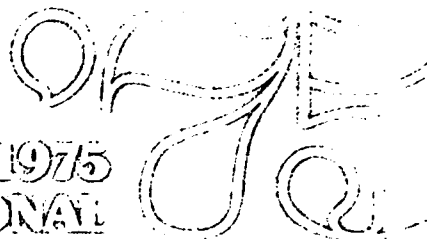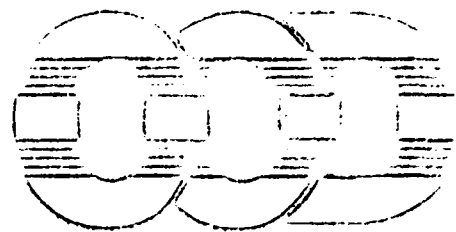
1975 INTERNATIONAL CONFERENCE ON THE APPLICATION OF CHARGE-COUPLED DEVICES

PROCEEDINGS

THE USE OF CHARGE COUPLED DEVICES IN ELECTROOPTICAL PROCESSING

M. A. Monahan, R. P. Bocker, K. Bromley, A. C. H. Louie,
R. D. Martin, and R. G. Shepard

U.S. Naval Electronics Laboratory Center
Electrooptics and Optics Division
San Diego, California

ABSTRACT. The use of coherent optical analog methods to perform matrix multiplications has been reported in the literature. Described in this paper are two incoherent optical systems in which the matrix multiplication represents a general linear transformation of one data vector into another. The input data vector is represented as a time sequence of N amplitude weighted electrical pulses which temporally modulate the light output of an LED. This modulated light field first passes through optical transparency and is then incident upon a charge coupled device (CCD). In one system the CCD is a 500 x 1 element line array and in the other it is a 100 x 100 area array. The transparency is arranged in a rectangular array of M x N elements, where the transmittance of each element is proportional to the m, n^{th} sample of the impulse response or kernel of the linear transformation. Finally, the output data vector is in the form of a time sequence of M electrical pulses, the amplitudes of which represent the values of the desired output data vector.

The linear transformation thus performed is one of broad application, the particular nature of the transformation depending upon the form of the impulse response encoded into the optical memory transparency. Examples of transformations which can be readily programmed into the device include convolutions; correlations; Fourier, Laplace, and Walsh-Hadamard transformations; and linear filtering.

An electrooptical system is particularly appropriate for such calculations in moderate-accuracy (6-8 bit) applications. A direct evaluation of this discrete matrix operation requires M x N analog multiplications to be performed in a sequential manner such that the elements of the output vector are produced one at a time. To significantly reduce the processing time relative to such a slow direct implementation, one must reduce the time required for each analog multiplication, process in parallel, or both. In the electrooptical systems described here, both of these processing advantages are inherently present.

INTRODUCTION

The use of coherent optical analog methods to perform matrix multiplications has been reported in the literature.[1-3] An electrooptical approach utilizing incoherent optical technology has recently been described in which a vidicon tube was used as the integrating element.[4-6] Presented below is an extension of this latter work in which the vidicon is replaced by a line-array charge coupled device (CCD) in one system, and by an area-array CCD in another.[7, 8]

We consider the electrooptical implementation of a general linear filter, illustrated in Fig. 1, which is characterized by an impulse response h(u,v). The output of such a filter, g(v), is related to the input, f(u), through a general linear transformation

$$\int_{-\infty}^{\infty} f(u)h(u;v)du = g(v) \qquad (1)$$

in which the impulse response appears as a weighting function. We shall not restrict our discussion to shift-invariant filters, for which Eq. (1) would reduce to a simple convolution, and will therefore be able to treat a larger class of useful linear transformations. The relation described by Eq. (1) is one of broad application, the particular nature of the transformation depending on the form of the impulse response. Table 1 lists a few examples of signal processing transformations together with the appropriate form of h(u,v). Thus, a signal processing device for which the impulse response can be readily programmed, and which can subsequently perform the transformation of Eq. (1) with economy

Attachment 12

217

of time and hardware, should find widespread and versatile use.
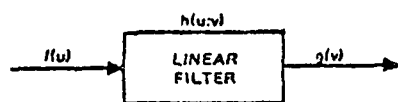
Fig. 1. General linear filter characterized by impulse h(u;v). Input f(u) is transformed into output g(v) through a linear transformation integral.

TABLE 1. EXAMPLES OF LINEAR TRANS-FORMATIONS COMMONLY USED IN SIGNAL PROCESSING APPLICATIONS. EACH TRANS-FORMATION IS OF THE FORM OF EQ. (1), WITH IMPULSE RESPONSE h(u;v) AS SHOWN BELOW.

TRANSFORMATION	IMPULSE RESPONSE
Convolution	$h(v-u)$
Cross correlation	$h(u-v)$
Autocorrelation	$f(u-v)$
Cosine transform	$\cos(2\pi uv)$
Fourier transform	$\exp(-i2\pi uv)$
Laplace transform	$\exp(-uv)$
Hankel transform	$2\pi J_0(2\pi uv)u$

Anticipating the electrooptical implementations to be described below, which are basically sampled data systems, we consider the discrete finite version of Eq. (1).

$$\sum_{n=0}^{N-1} f_n h_{mn} = g_m \quad m = 0,1,2,\ldots,M-1 \quad (2)$$

It is often useful to rewrite this equation in its equivalent matrix notation

$$[H][F] = [G].$$

Written in full this relation becomes

$$(3)$$

where each matrix has a direct interpretation in terms of the processors described in this paper.

At this point, it is appropriate to comment on why an electrooptical implementation of this fundamental matrix operation is being considered here. Note that a direct evaluation of Eq. (3) requires $M \times N$ analog multiplications to be performed in a sequential manner such that output values g_m are produced one at a time. In order to increase the rather slow processing rate associated with such a direct approach, one must reduce the time required for each multiplication, process in parallel, or both. In an optical system both of these processing advantages are inherently present. Although they must still be detected, analog multiplications take place as fast as light travels through an optical transparency (about 10^{-13} sec). Also, the two-dimensional nature of image transfer in an optical system provides the capability of performing many such multiplications simultaneously (up to about 10^6 in the systems described below). Therefore, in applications involving high-speed calculations of moderate accuracy (6-8 bit), an electrooptical system seems a particularly appropriate approach.

We describe in sections to follow two different electrooptical devices designed to implement the matrix multiply operation of Eq. (3). The first utilizes a scanning mirror to sweep the temporally modulated image of an optical memory transparency or mask across a line-array CCD. The second eliminates the need for a scanning mirror by replacing the line-array detector with an area-array CCD and electronically scanning the mask image within the detector itself.

LINE-ARRAY PROCESSOR

GENERAL DESCRIPTION

Fig. 2 depicts the incoherent electrooptical system used to evaluate the matrix multiplication of Eq. (3). The system consists of: (a) light emitting diode (LED), (b) condensing lens, (c) optical memory mask, (d) imaging lens, (e) scanning mirror, and (f) line-array CCD.

Given a temporal signal f(t), for which some linear transformation must be performed according to Eq. (1), the input to the device is a time sequence of electrical pulses, f_n, which represent sampled values of f(t). These samples are proportional to the elements of the column vector [F] in Eq. (3), and appear as an intensity modulation of the LED. The condensing lens is chosen for uniformity of illumination upon the mask and to maximize the light throughput in the system by imaging the light source into the entrance pupil of the imaging lens. Directly behind the condenser is placed the optical mask in which is encoded the matrix operator

218

[H]. An image of the mask is then formed by the imaging lens, via a scanning mirror, on the face of the CCD. The scanning mirror is galvanometer driven in a sawtooth fashion such that an image of the mask is repetitively swept across the CCD face at a constant velocity in a direction perpendicular to the long dimension of the array. The CCD is then allowed to integrate the light falling on it during a single passage of the image, and a new output vector [G] is generated and clocked out of the CCD at the end of each minor sweep.
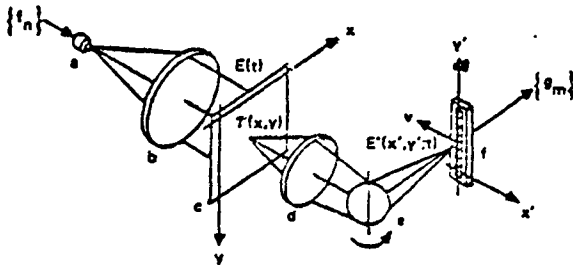
Figure 2. Incoherent electrooptical processor utilizing a mirror scan and line-array CCD. Components are: (a) light emitting diode (LED); (b) condensing lens; (c) optical memory mask; (d) imaging lens; (e) scanning mirror; and (f) line-array charge coupled device (CCD).

MATHEMATICAL ANALYSIS

Given that an analog temporal signal $f(t)$ must be sampled and then transformed according to Eq. (2) into a new discrete signal, we shall proceed by tracing $f(t)$ through the system from input to output. The first step in preparing the analog input signal for processing is to convert it to a time sequence of pulses by passing it through a form of sample-and-hold circuit. The discrete version of the signal, shown in Fig. 3, then becomes

$$f_s(t) = \sum_{n=0}^{N-1} f_n \, \text{rect}\left(\frac{t-nT-d/2}{d}\right) \qquad (4)$$

where

$$f_n \triangleq f(nT)$$

and T is the sampling period, d is the constant pulse duration, and the rectangle function (rect) is defined in Appendix A. The discrete signal $f_s(t)$ is then used to modulate the light emitted by the LED so that the spatially uniform irradiance distribution incident on the optical memory mask in the x,y plane is

$$E(t) = c_1 f_s(t) \qquad (5)$$

where the constant c_1 is a scaling factor which depends on the design of the condensing optics and on the scale of the electrical-to-optical pulse conversion by the LED and its electronics.
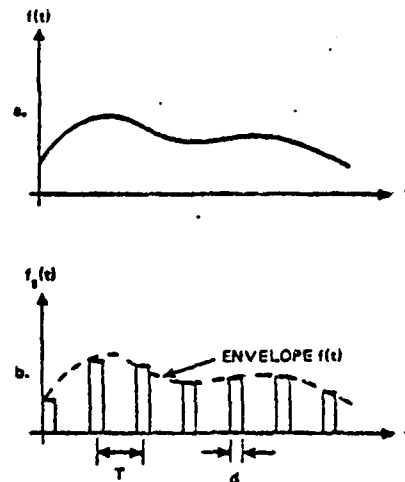
Fig. 3. The input signal (a) and its sampled version (b). Sample pulses of height $f_n \triangleq f(nT)$ and period T are all of same duration d.

The mask itself consists of a total of $M \times N$ rectangular apertures arranged in a rectangular array as shown in Fig. 4. Note that each element is the same size, $A \times B$, but that the clear area of each element, $a_{mk} \times W$, is modulated by varying the x-dimension while holding the y-dimension fixed. This design, which modifies the transmitted radiant flux by varying the area of an aperture, is considerably easier to fabricate than one which varies the transmission properties of some photographic or electrooptic material. From the diagram, then, the transmittance function of the optical mask is given by

$$\tau(x, y) = \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} \text{rect}\left(\frac{x-kA}{a_{mk}}\right)\text{rect}\left(\frac{y-mB}{W}\right) \qquad (6)$$

where

$$a_{mk} = c_2 h_{mk} \qquad (7)$$

and c_2 is a scaling constant. Inspection of Eq. (6) shows that there is now a one-to-one correspondence between each element of the mask and the corresponding element of the matrix operator [H] in Eq. (3).

Attachment 12

219

END

FILMED

DTIC
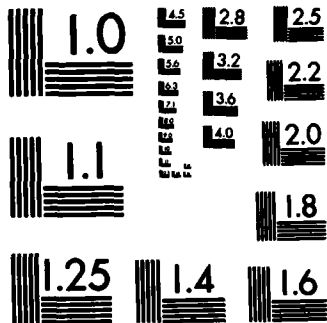
MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A
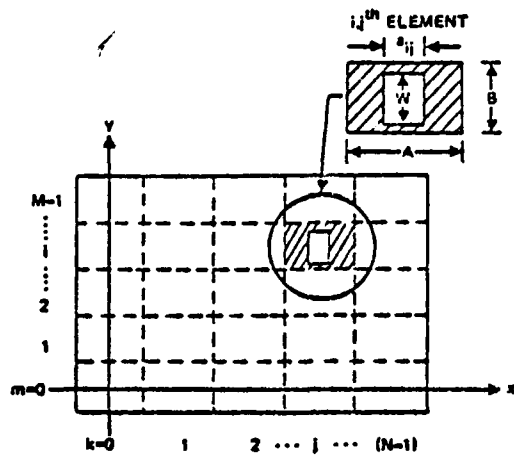
Fig. 4. Optical memory mask showing only i,j^{th} element. Mask consists of $M \times N$ rectangular elements, the clear portion of each having the same width W but a variable length a_{mk}.

The light field emerging from the optical mask is simply the product of the incident irradiance and the mask transmittance function. The resulting light distribution is then imaged onto the plane of the detector via a scanning mirror. The irradiance distribution in the x', y' plane resulting from the temporally modulated mask image, moving at a speed v in the negative x' direction, is then

$$E'(x',y';t) = c_3 E(t)\tau\left(\frac{x'+vt-x'_0}{\beta}, \frac{y'}{\beta}\right) \qquad (8)$$

In this equation c_3 is a constant determined from radiometric considerations of the imaging system, x'_0 is an arbitrary spatial phase term associated with the scanning, and β is the lateral magnification associated with the mapping of the optical mask into its image.

Knowing the irradiance distribution in the detector plane, the quantity of interest in terms of the response of the CCD is the total exposure delivered to the x',y' plane during a single sweep of the mirror. This is given by

$$\xi(x',y') = \int_0^\infty E'(x',y';t)dt. \qquad (9)$$

Making the appropriate substitutions from Eq. (4) through (8), the total exposure becomes

$$\xi(x',y') = c_1 c_3 \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} f_n \text{ rect}\left(\frac{y'-mB'}{W'}\right)$$

$$(10)$$

$$\int_0^\infty \text{rect}\left[\frac{t-(nT+d/2)}{d}\right] \text{rect}\left[\frac{t-(kA'/v+x_0/v)+x'/v}{a'_{mk}/v}\right] dt$$

where for convenience we have defined the magnification-scaled quantities

$$A' = \beta A$$
$$B' = \beta B$$
$$W' = \beta W$$
$$a'_{mk} = \beta a_{mk}$$

Evaluating the integral in Eq. (10) we find that the m,k^{th} exposure element due to the nth LED pulse is one of three possible trapezoidal solids.[*]

$$\xi(x',y')_{nmk} = \begin{cases} c_1 c_3 df_n \text{rect}\left(\frac{y'-mB'}{W'}\right) \text{trap}\left[\frac{x'-k-n1A'}{\frac{a'_{mk}+d}{2};\frac{a'_{mk}-vd}{2}}\right] & a'_{mk} > vd. \quad (11a) \\[2em] \frac{c_1 c_3}{v} f_n a'_{mk} \text{rect}\left(\frac{y'-mB'}{W'}\right) \text{tri}\left[\frac{x'-k-n1A'}{a'_{mk}}\right] & a'_{mk} = vd. \quad (11b) \\[2em] \frac{c_1 c_3}{v} f_n a'_{mk} \text{rect}\left(\frac{y'-mB'}{W'}\right) \text{trap}\left[\frac{x'-k-n1A'}{\frac{vd+a'_{mk}}{2};\frac{vd-a'_{mk}}{2}}\right] & a'_{mk} < vd. \quad (11c) \end{cases}$$

In arriving at the expressions of Eq. (11), several important requirements have been included. First, it is assumed that the moving mask image is in phase synchronization with the LED such that the first ($k=0$) column of the mask image is centered on the y'-axis half way through the first ($n=0$) LED pulse. This implies that

$$x'_0 = \frac{vd}{2}.$$

Second, the timing of the LED pulses must be such that the exposure pattern will be of the proper scale. That is, it will ensure that half way through the nth light pulse the $k=n^{th}$ column of the mask image will also be centered on the y'-axis. This means that

$$A' = vT.$$

------

[*]See Appendix A for definitions of the rect, trap, and tri functions used here. Also see Appendix B for evaluation of the form of integral contained in Eq. (10).

Attachment 12

The exposure elements described by Eq. (11) and their size relationships to a CCD element are illustrated in Fig. 5. The exposure element $\xi(x',y')_{nmk}$ describes the distribution of radiant energy per unit area delivered to the detector plane from the m,k^{th} mask element during the nth LED pulse. As can be seen from the diagram, the size of a given exposure element relative to the size of a CCD element must be considered when computing the total radiant energy actually entering the CCD element. Of the five possible combinations described in Fig. 5, three pass an amount of radiant energy during a single light pulse which is proportional to the desired product $f_n a'_{mk}$. These are cases (a) and (b) with trapezoid height $R = (c_1 c_3 / v) f_n a'_{mk}$ and case (b) with $R = c_1 c_3 d f_n$. Of these three we select the first for practical consideration since it greatly reduces tolerance requirements in mask fabrication and in scan synchronization. Therefore, combining Eq. (10) and (11c), the total exposure in the x',y' plane is

$$\xi(x',y') = \frac{c_1 c_3}{v} \sum_{n=0}^{N-1} \left\{ \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} f_n a'_{mk} \right. \tag{12}$$

$$\left. \text{rect}\left(\frac{y'-mB'}{W'}\right) \text{trap}\left[\frac{x'-(k-n)A'}{\frac{vd-a'_{mk}}{2}; \frac{vd+a'_{mk}}{2}}\right] \right\}$$
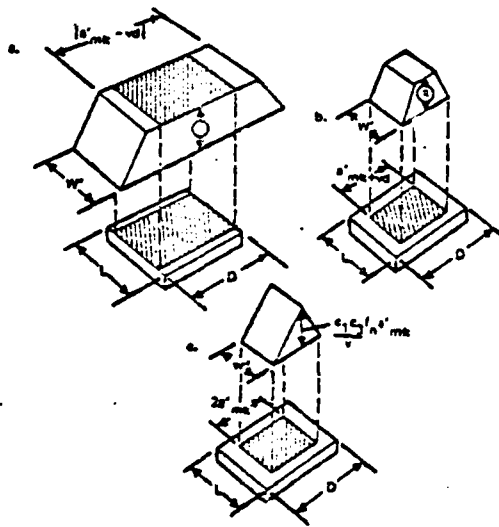
Fig. 5. Possible relationships between exposure element and detector element as described by Eq. (11). Eq. (11a) can correspond to cases (a) or (b) above where $R = c_1 c_3 d f_n$. Eq. (11b) is represented by case (c) above, and Eq. (11c) is described by cases (a) or (b) with $R = c_1 c_3 / v \, f_n a'_{mk}$.

As shown in Fig. 6, the expression in braces, $\left\{ \quad \right\}$, above, represents a rectangular array of trapezoidal solids due to the nth LED pulse. A new such exposure array is created with each input pulse f_n, but each array is shifted from the preceding by an amount A'. Note from the figure that for the nth input light pulse, only the $k=n^{th}$ column of the exposure array will be superimposed upon the detector. Finally, we assume that in the y'-dimension, each exposure element is narrower than a CCD element. In summary, the assumptions

$$k = n$$
$$L \geqslant W'$$
$$D \leqslant vd - a'_{mn}.$$

when combined with Eq. (12), yield the total radiant energy entering the mth CCD element during a complete mirror sweep as

$$Q_m = c_1 c_2 c_3 \frac{LD}{v} \sum_{n=0}^{N-1} f_n h_{mn}. \tag{13}$$
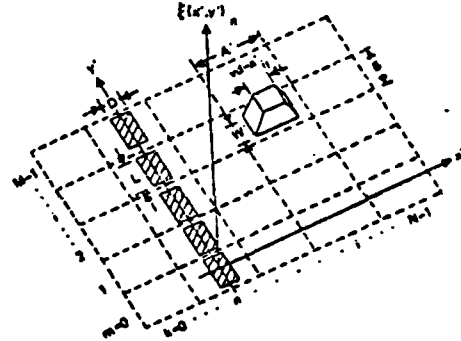
Fig. 6. Exposure pattern $\xi(x',y')_n$ due to the nth LED pulse. Pattern is rectangular array of trapezoidal solids (only i,jth exposure element shown) shifted in the negative x' direction by an amount nA'. Note that n^{th} column of exposure array is centered on the detector array (shaded elements).

Comparison of Eq. (13) with Eq. (2) shows that the quantity Q_m is indeed proportional to the desired quantity g_m. At the completion of a given mirror sweep the charge packets stored in each of the M CCD elements are clocked sequentially out of the device yielding a time sequence of pulses which are respectively proportional to the elements of the desired column vector [G] in Eq. (3).*

---

*The assumption is that the energy entering each CCD element is linearly converted to charge within the element. Deviations from this assumption, or compensation techniques, will not be discussed in this paper.

Attachment 12

One final consideration is the relation between the length of the mask elements, a_{mk}, and their spacing in the x-dimension. A profile of the m,n^{th} trapezoid function of Eq. (12) is shown in Fig. 7, where it is superimposed upon the mth CCD element of width D.

For a given pulse duration d and velocity v, note from the figure that, while the scope of the trapezoid sides depends upon the LED pulse strength f_n, the overall width of the exposure element depends solely upon the length of the mask image element a'_{mn}. Also notice that the upper vertices of the trapezoid lie on the legs of a triangle shown by dashed lines. Thus, for a given f_n, the top of the trapezoid gets higher and narrower as a'_{mn} increases. Since the detector must lie within the flat portion of the exposure element, a'_{mn} can vary between zero and a maximum value determined by the CCD element width D. In terms of the mask element itself,

$$0 \leqslant a_{mn} \leqslant \frac{vd - d}{\beta} . \tag{14}$$
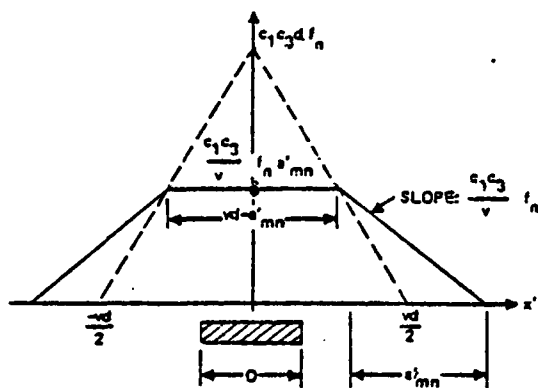
Fig. 7. Profile of m,n^{th} exposure element superimposed upon m^{th} CCD element (shaded). Uper vertices of trapezoid always lie on legs of triangle shown with dashed lines.

To calculate the minimum spacing required between mask elements, the overlap between two adjacent trapezoidal elements for which a'_{mn} is a maximum must be considered. As shown in Fig. 8, the minimum allowable spacing in the exposure plane then occurs when d=T. From the figure we conclude that the absolute minimum spacing of mask elements must therefore be

$$A_{min} = \frac{vT}{\beta} . \tag{15}$$
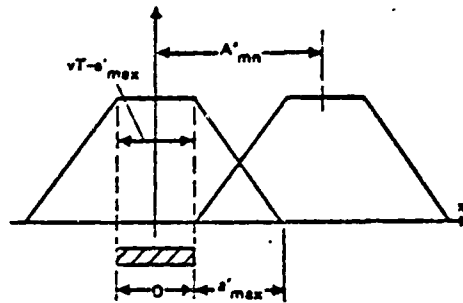
Fig. 8. Minimum possible spacing between adjacent exposure elements occurs when $d = T$ and when a'_{mn} assumes the maximum value $a'_{max} = vT - D$.

AREA-ARRAY PROCESSOR

GENERAL DESCRIPTION

In the system described below, the need for a scanning mirror is eliminated by incorporating an area-array detector in place of the line-array CCD used above. By using a two-dimensional CCD, the scanning of the mask image can now be performed electronically within the detector itself. This allows considerable simplification in system design as well as analysis. Such a modified system is represented in Fig. 9, where up to the optical memory mask the geometry is essentially the same as in the line-array system described by Fig. 2. Immediately behind the mask is an area-array CCD whose output is a sequence of pulses representing the desired vector [G]. This geometry not only avoids the mechanical complexity of a scanning mirror, but also eliminates the imaging lens and the space associated with its mapping of the mask image onto the detector.[*]
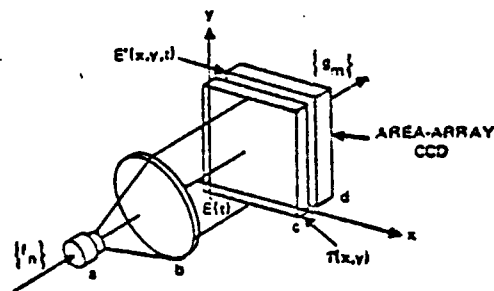
Fig. 9. Area-array electrooptical processor in which need for a scanning mirror is eliminated through use of a two-dimensional CCD. System consists of: (a) LED; (b) condensing lens; (c) optical memory mask; and (d) area-array CCD.

[*]We speak here of the mask and detector as being in physical contact, as indeed they could be with a specially designed CCD. However, for convenience in the experimental work performed thus far, the mask has been imaged onto the detector with a lens.

Attachment 12

MATHEMATICAL ANALYSIS

As before, we represent the spatially uniform light field incident upon the optical mask by

$$E(t) = c_1 \sum_{n=0}^{N-1} f_n \operatorname{rect}\left(\frac{t-nT-d/2}{d}\right) \qquad (16)$$

where c_1 is a constant scale factor.

For ease of fabrication, we again represent the elements of the optical mask as a rectangular clear apertures arranged on a rectangular array with spacing $A \times B$ (Fig. 10). Here the clear portion of the m,k^{th} mask element is of length $\sqrt{a_{mk}}$ in the x-dimension and width $K\sqrt{a_{mk}}$ in the y-dimension where

$$a_{mk} = c_2 h_{mk}, \qquad (17)$$

$$K = B/A, \qquad (18)$$

and c_2 is a scaling constant.

The transmittance function of the mask is then

$$\tau(x,y) = \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} \operatorname{rect}\left(\frac{x-kA}{\sqrt{a_{mk}}}\right) \operatorname{rect}\left(\frac{y-mB}{K\sqrt{a_{mk}}}\right) \quad (19)$$
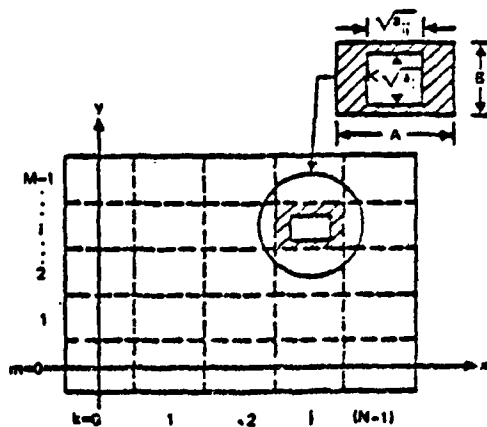
Fig. 10. Optical memory mask showing only i,jth element. Mask consists of rectangular array of M x N rectangular elements. The clear area of the i,jth element is Ka_{ij}, where $K = B/A$.

Immediately behind the optical mask the irradiance distribution incident on the detector plane is

$$E'(x,y;t) = E(t)\tau(x,y). \qquad (20)$$

The optically sensitive region of the area-array CCD consists of a rectangular array of identical rectangular photosensors, as shown in Fig. 11. These CCD elements, each of size $D \times L$, are arranged on an array the same size and scale as the optical mask. In addition, the aspect ratio of each mask element is scaled to be the same as that of the corresponding CCD element. That is,

$$\frac{B}{A} = \frac{L}{D} = K. \qquad (21)$$
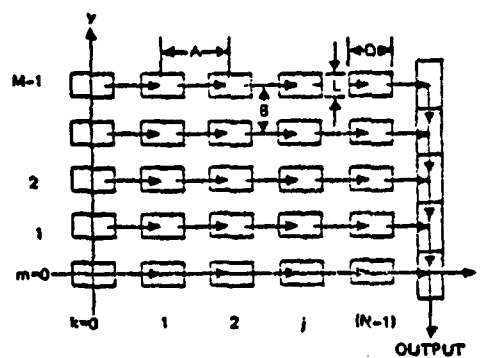
Fig. 11. Rectangular array of rectangular CCD photosensors. Each element is of size D x L, where $L/D = B/A = K$. Column at right represents output shift register for clocking out charge transferred from the photosensor array.

We now focus attention on the energy entering the m,k^{th} element of the CCD due to the nth LED pulse,

$$Q_{nmk} = \int\int\int_{-\infty}^{\infty} E'(x,y;t)dxdydt$$

$$= c_1 c_2 dK f_n h_{mk}. \qquad (22)$$

In particular, the energy entering the $k = n^{th}$ element in the mth row is proportional to $f_n h_{mn}$, the product inside the summation of Eq. (2). What is desired is the sum of such products for all values of n. Thus, if the charge content of each CCD cell in the mth row is transferred laterally, as indicated in Fig 11, by one element between LED pulses, then the stored photocharge in the m,k^{th} element due to the nth pulse can be added to the charge stored in the $(m,k-1)^{st}$ element due to the $(n-1)^{st}$ pulse. This means that the partial sum in the last $(k=N-1)$ cell of the mth row after the rth pulse is

$$S_{mr} = c_1 c_2 dK \sum_{n=0}^{r} f_n h_{m,N-(r-n+1)} \qquad (23)$$

Finally, after the last pulse ($r=N-1$), the charge stored in the last cell of the mth row is proportional to

$$S_{m,N-1} = c_1 c_2 dK \sum_{n=0}^{N-1} f_n h_{mn}. \qquad (24)$$

Again referring to Eq. (2), this expression is proportional to the desired quantity g_m. After N light pulses, the charge stored in the output shift register is vertically clocked out. This charge, in the form of discrete packets, yields a time sequence of pulses proportional to the values g_m of the desired column vector [G].

Fig. 12. Line-array electrooptical processor on a 4 x 5 inch circuit card. Device is programmable by inserting a desired program mask. Support electronics occupies three additional cards of the same size.

COMMENTS AND CONCLUSIONS

Eq. (13) and (24) predict that the line-array and area-array systems described above are capable of performing the linear transformation of ...q. (2). A developmental model of the line array processor, shown in Fig. 12, has been assembled and tested. The system is similar to that previously reported,[4-6] however, here a 500-element Fairchild line-array CCD has replaced the vidicon tube. In addition, a compact layout has been used in which the optical system is confined to a 4 x 5 inch circuit board.

Fig. 13. Area-array electrooptical processor. Device is programmable by inserting desired program mask.

An area-array processor which utilizes a 100 x 100 element Fairchild CCD has also been assembled, see Figure 13, and is currently being tested. The detector, which is a standard imaging chip, is being driven in a manner to suit this signal processing application. Used as an image sensor, the CCD array would continuously integrate during each full video frame.* However, for the case at hand the device integrates during each individual LED pulse, but between pulses the charge collected at each photosite is transferred laterally by one element and added to the photocharge from the next pulse. This shift-and-add process is repeated 100 times, after which the charge deposited in the output shift register, representing the desired output data vector, is clocked out.

As described earlier, the optical memory masks fabricated thus far utilize an area modulation scheme for encoding the values h_{mn}. This technique is straightforward, not involving materials problems (e.g., nonlinearities), and lends itself to mask fabrication with a programmable desk calculator and x-y plotter. Two mask examples are shown in Fig. 14 for the cases of a discrete identity transform and a discrete cosine transform.

---

*Interlacing is ignored in this discussion.

Attachment 12
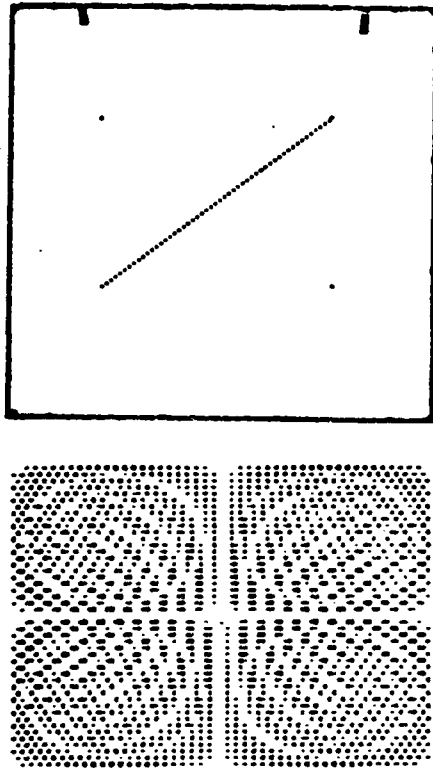
Fig. 14. Examples of 35 mm format memory masks designed to perform (a) a discrete identity transform, and (b) a discrete cosine transform.

These masks have been used in preliminary tests of the area-array processor, with favorable results illustrated in Figs. (15) and (16). As a first test we consider the case where the impulse response operator of Eq. (2) is a matrix with diagonal elements of unity and off-diagonal elements zero. This defines the so-called identity matrix. Its use in Eq. (2) reproduces at the output an exact replica of the input function. Typical performance of the area-array processor with an identity matrix in place is shown in Figure 15. In this example the input signal (lower trace) was a one volt (peak-to-peak) triangle wave of 0.3 kHz frequency, sampled at 10 kHz. As seen in the figure, the output signal (upper trace) is a well formed triangle wave of the same frequency as the input signal. The few spurious samples at each end of the output trace are not part of the processed signal and are ignored. As a second test, a mask was prepared for performing a discrete cosine transform. Theoretically, a pure cosine wave input results in two delta functions at the output, centered in the output array, and separated by a distance pro-

portional to twice the input frequency. The results of this test are shown in Fig. 16 for two input cosine waves of one volt (peak-to-peak) amplitude and frequencies (a) 1.1 kHz and (b) 2.5 kHz. As can be seen from the figure, these results agree quite well within the theoretically predicted outputs.

It is appropriate at this point to consider the data rates associated with these processors. In the 500 x 1 line-array system of Fig. 12, a rather slow scanning mirror was used rather than a high speed spinning prism to demonstrate the concept. The data must be emptied out of the CCD at the end of each mirror scan before a new scan can begin. The time required to perform the transform on the next set of input samples is determined by the 20 msec sweep period of the mirror. In this system, input samples can be fed in at a continuous rate of about 1 kHz. The resulting output comes in 0.5 msec bursts of 500 data pulses at 1 MHz with 20 msec between bursts.

For the 100 x 100 area-array processor, the data can also be clocked from the output shift register at 1 MHz. This must be 100 times faster than the rate at which charge is being transferred across the array. Since for each lateral data shift there is one light pulse, a continuous input rate of 10 kHz yields a continuous output pulse rate of 1 MHz.

Figure 15. Input (lower trace) vs. output (upper trace) with identity matrix programmed into electrooptical processor.

Attachment 12
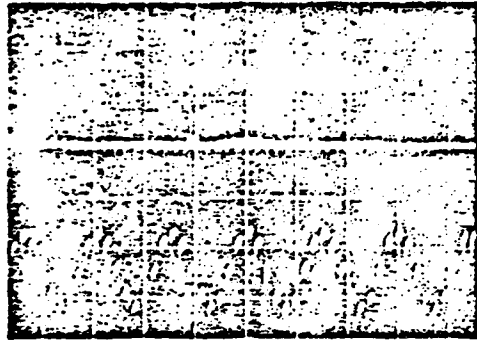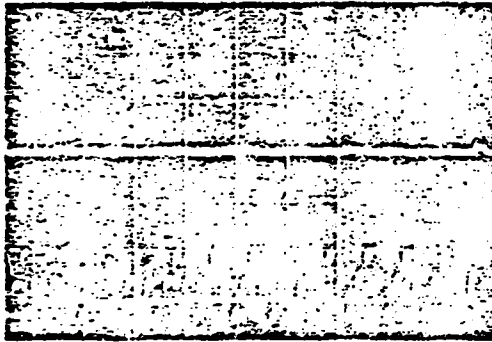
Figure 16. Input (lower trace) vs. output (upper trace) with processor programmed for a cosine transform. Input signals are cosine waves with frequencies (a) 1.1 kHz and (b) 2.5 kHz.

In performing a linear transformation, the line-array device operates on sequential windows of input data as shown in Fig. 17a. Although the output data appear in high-frequency bursts, over the period of one mirror cycle there are 500 output data pulses for each 500 input samples. However, in this area-array processor example there are 100 output data pulses for each input sample. What this means is that a new and complete transform is computed, with each input pulse, on a sliding window of input data (Fig. 17b). That is, each new set of output pulses represents the linear transformation of an input data set which differs from the preceding set by the addition of a new sample and the dropping of the oldest. If the functional form of the input signal varies in time, then its transform as a function of time can be continuously computed. This represents a useful capability inherently available in the device if required. An example of its use would be the computation of the discrete Fourier transform of a signal whose exact time of arrival is not known. Thus, one could avoid truncating the input signal by not including it entirely within the sampling window.

As a final comment, let us point out that although we have considered the analog input f(t) to have been sampled before modulating the LED, this is not in general necessary. When the input signal is sampled according to Eq. (4), the total light collected by a given CCD element due to the nth LED pulse is exactly proportional to f_n. However, if the analog input is not sampled beforehand, then the "sampling" is done in effect by the CCD itself. The photosensitive elements of the CCD integrate the light incident upon them only during a time determined by the length of a photogate clocking pulse. This performs a sampling operation. However, the light integrated will be proportional to the average value of f(t) during the sampling interval. This average value will in most cases be sufficiently close to the instantaneous value f_n so that sampling of the input signal and all the associated synchronization problems can be eliminated.
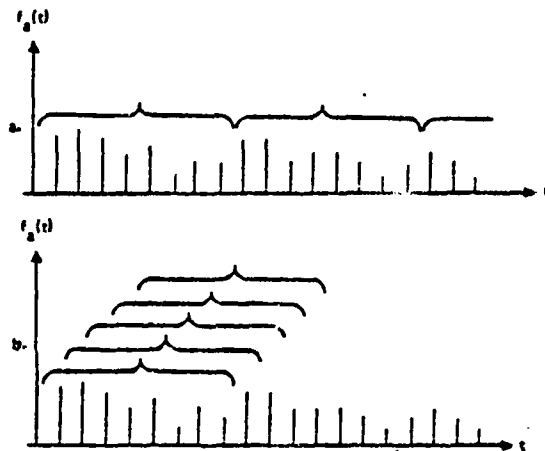
Fig. 17. With the area-array processor, transforms can be performed not only on (a) sequential windows of data but also on (b) input data under a sliding window.

ACKNOWLEDGMENT

REFERENCES

1. R. A. Heinz, J. O. Artman, and S. H. Lee, Appl. Opt., 9,2161 (1970).

2. D. P. Jablonowski, R. A. Heinz, and J. O. Artman, Appl. Opt., 11,174 (1972).

3. L. J. Cutrona, Optical and Electro-Optical Information Processing, J. T. Tippet et al., Eds. (MIT Press, Cambridge, 1965), p. 97-98.

Attachment 12

4. R. P. Bocker, Applied Optics, 13, 1670 (1974).

5. K. Bromley, Optica Acta, 21, 35 (1974).

6. R. P. Bocker, Ph.D. Dissertation, University of Arizona, June 1975.

7. R. P. Bocker, K. Bromley, and M. A. Monahan, "Optical Data Processing for Fleet Applications," *Naval Research Reviews*, (Office of Naval Research, Arlington, VA), p. 44, May – June 1974.

8. M. A. Monahan, R. P. Bocker, K. Bromley, and A. Louie, "Incoherent Electrooptical Processing with CCDs," *International Optical Computing Conference April 23–25, 1975, Digest of Papers*, (IEEE Catalog No. 75 CH0941-5C), p. 25

APPENDIX A

Many useful and often common functions must be defined in piecewise fashion because of abrupt changes in the value of the function. For example, consider the function $f(u)$ such that

$$f(u) = \begin{cases} 0, & u < -a, \\ \dfrac{u}{a}+1 & -a \leqslant u < 0, \\ -\dfrac{u}{a}+1, & 0 \leqslant u < a, \\ 0, & u > a. \end{cases}$$

To achieve compactness and clarity of notation for such simple but awkwardly expressed functions, we *define* in Fig. A1 a *set* of functions which implicitly include such abrupt behavior. We refer to these as the rectangle, triangle, and trapezoid functions, respectively. Note that the piecewise function cited in the above example now may be simply written as

$$f(u) = \text{tri}\left(\frac{u}{a}\right).$$

$$\text{rect}\left(\frac{u}{a}\right) \overset{\Delta}{=}$$

$$\text{tri}\left(\frac{u}{a}\right) \overset{\Delta}{=}$$

$$\text{trap}\left(\frac{u}{a \pm b}\right) \overset{\Delta}{=}$$
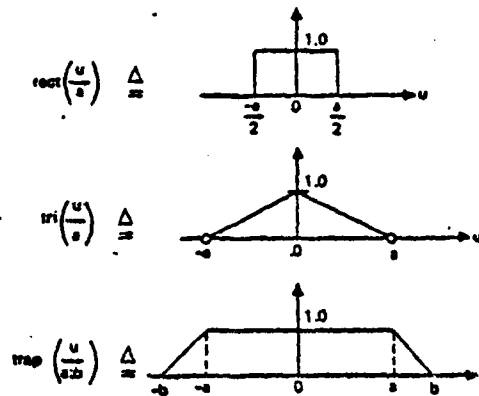
Fig. A1. Calculations involving abruptly changing functions are greatly simplified by adopting a compact notation. Used in this paper are three such awkwardly expressed functions which are simply defined as the (a) rectangle (rect), (b) triangle (tri), and (c) trapezoid (trap) functions, respectively.

APPENDIX B

The integral contained in Eq. (10) is of the form

$$\int_{-\infty}^{\infty} \text{rect}\left(\frac{u}{a}\right) \text{rect}\left(\frac{u \pm v}{b}\right) du. \qquad (B1)$$

The integrand in Eq. (B1) gives the area common to both rectangle functions when the second is offset from the first by an amount v (Fig. B1). Thus, the overlap area, which varies as a function of v, provides three possible solutions to the integral:

$$a \text{ trap}\left[\frac{v}{\left(\frac{b-a}{2}\right) ; \left(\frac{b+a}{2}\right)}\right], \qquad a < b. \qquad (B2a)$$

$$b \text{ tri}\left(\frac{v}{b}\right), \qquad a = b, \qquad (B2b)$$

$$b \text{ trap}\left[\frac{v}{\left(\frac{a-b}{2}\right) ; \left(\frac{a+b}{2}\right)}\right], \qquad a > b. \qquad (B2c)$$
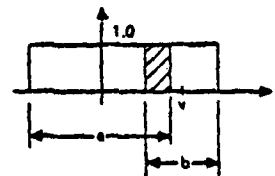
Fig. B1. Shaded overlap area between two rect functions gives value of integral in Eq. (B1) as function of displacement v.

THE MOUTH ORGAN

ee   i

ah   a

I   I

aw   ɔ

ae   ae

oo   u
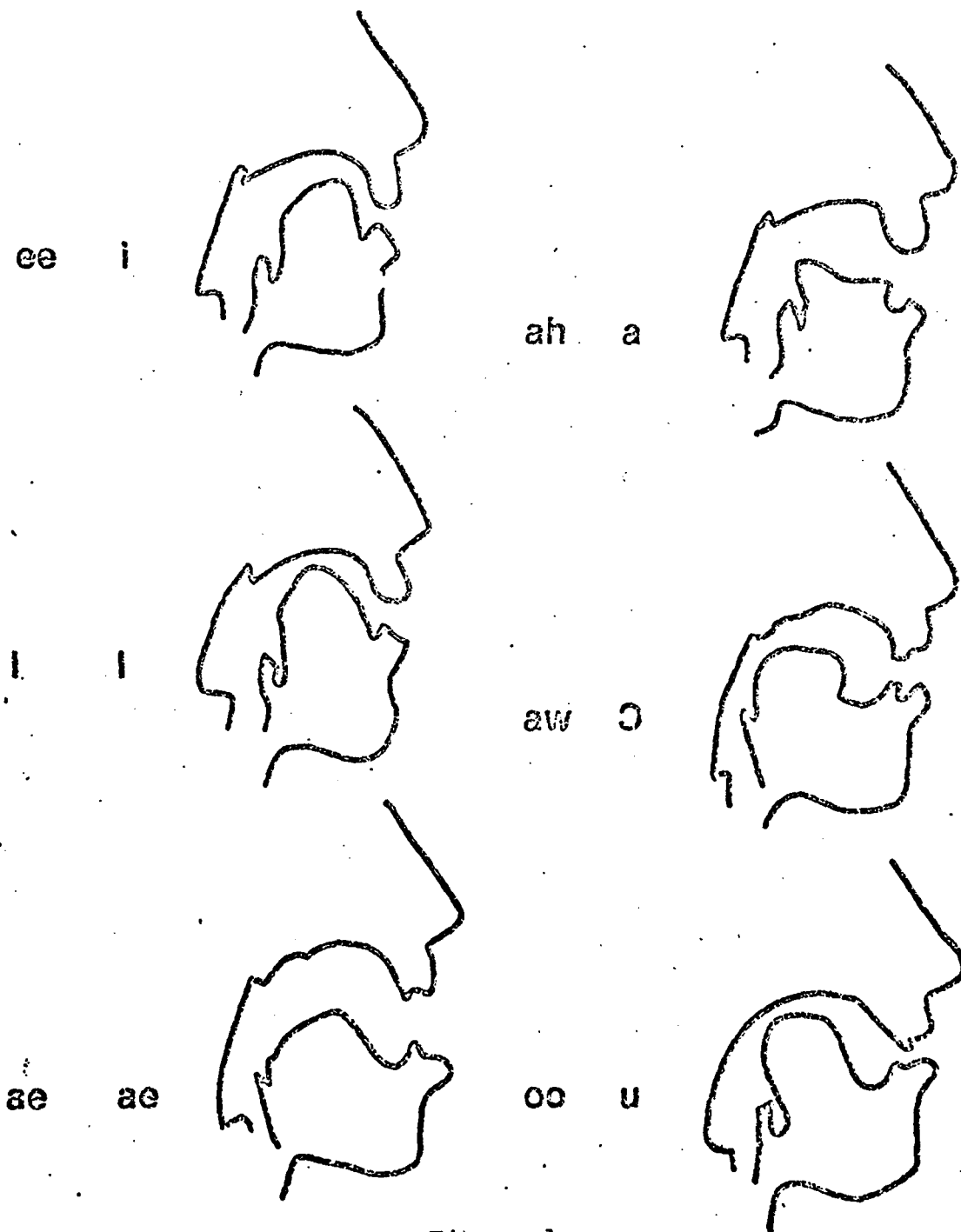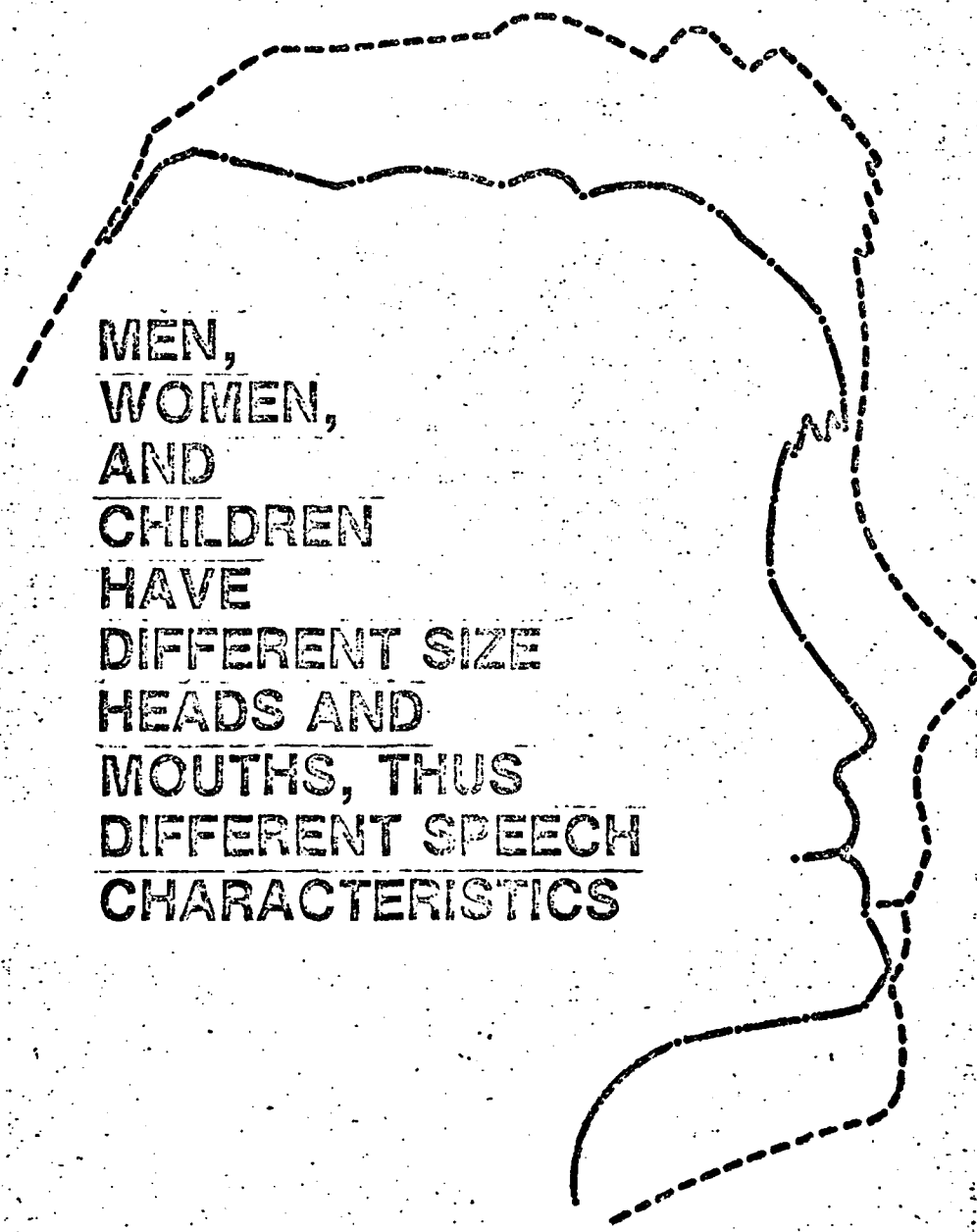
Figure 1

MEN,
WOMEN,
AND
CHILDREN
HAVE
DIFFERENT SIZE
HEADS AND
MOUTHS, THUS
DIFFERENT SPEECH
CHARACTERISTICS

Figure 2.

Attachment 13

2

VOWEL FREQUENCIES

LEGEND: ···· MEN   -- WOMEN   ·-·- CHILD

FREQUENCY - HERTZ

3RD FORMANT

3,000 —

2ND FORMANT

2,500 —

2,000 —

1,500 —

1,000 —

1ST FORMANT

500 —

VOCAL CORD

0 —

| ee | I | e | ae | ah | aw | u | oo | Λ | er |
| i | I | E | ae | a | ɔ | ʊ | μ | Λ | ɜ |
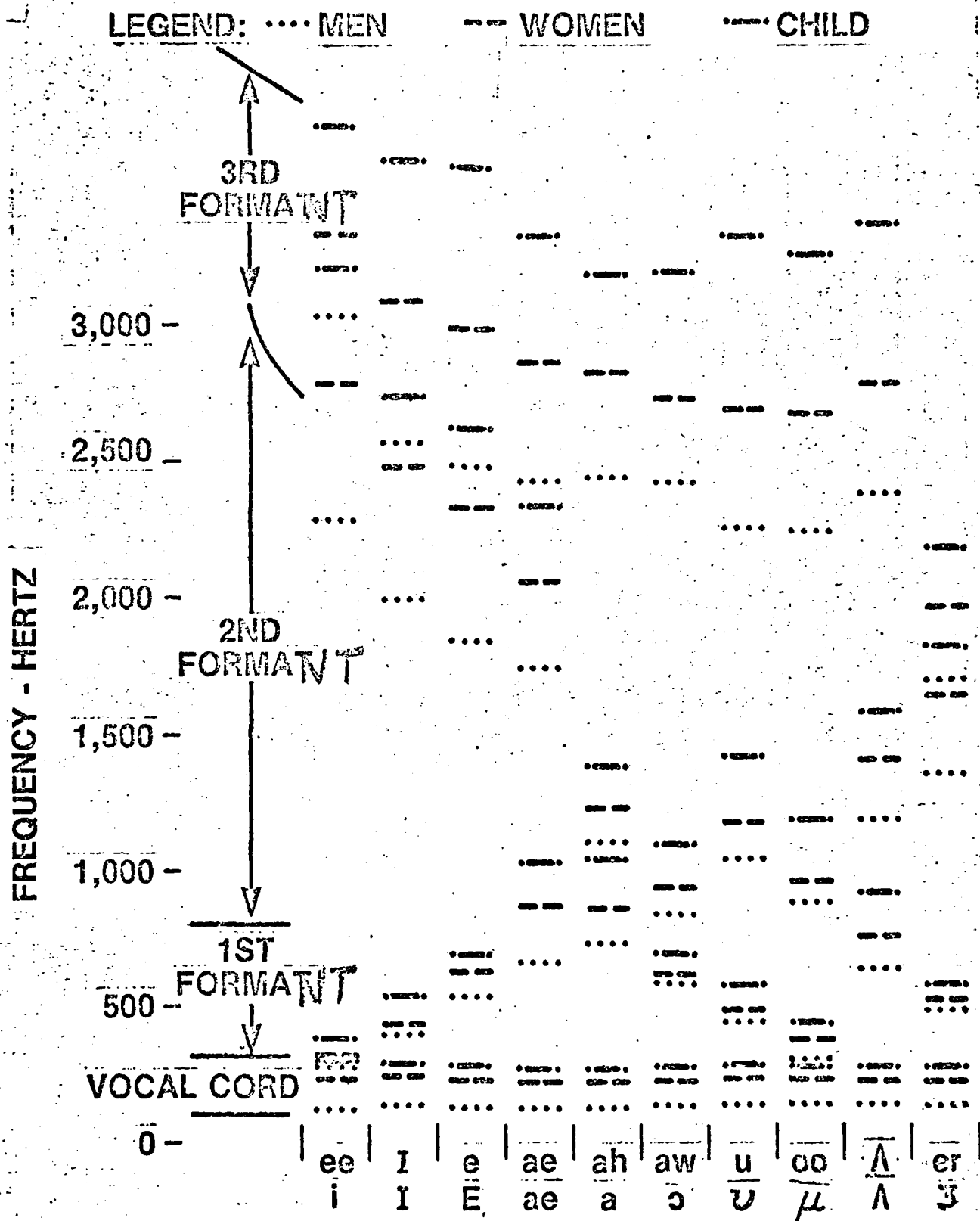
(BASED ON DATA FROM PETERSON AND BARNEY, 1951)

Figure 3

Attachment 13

3

VOWEL DIFFERENTIATIONS
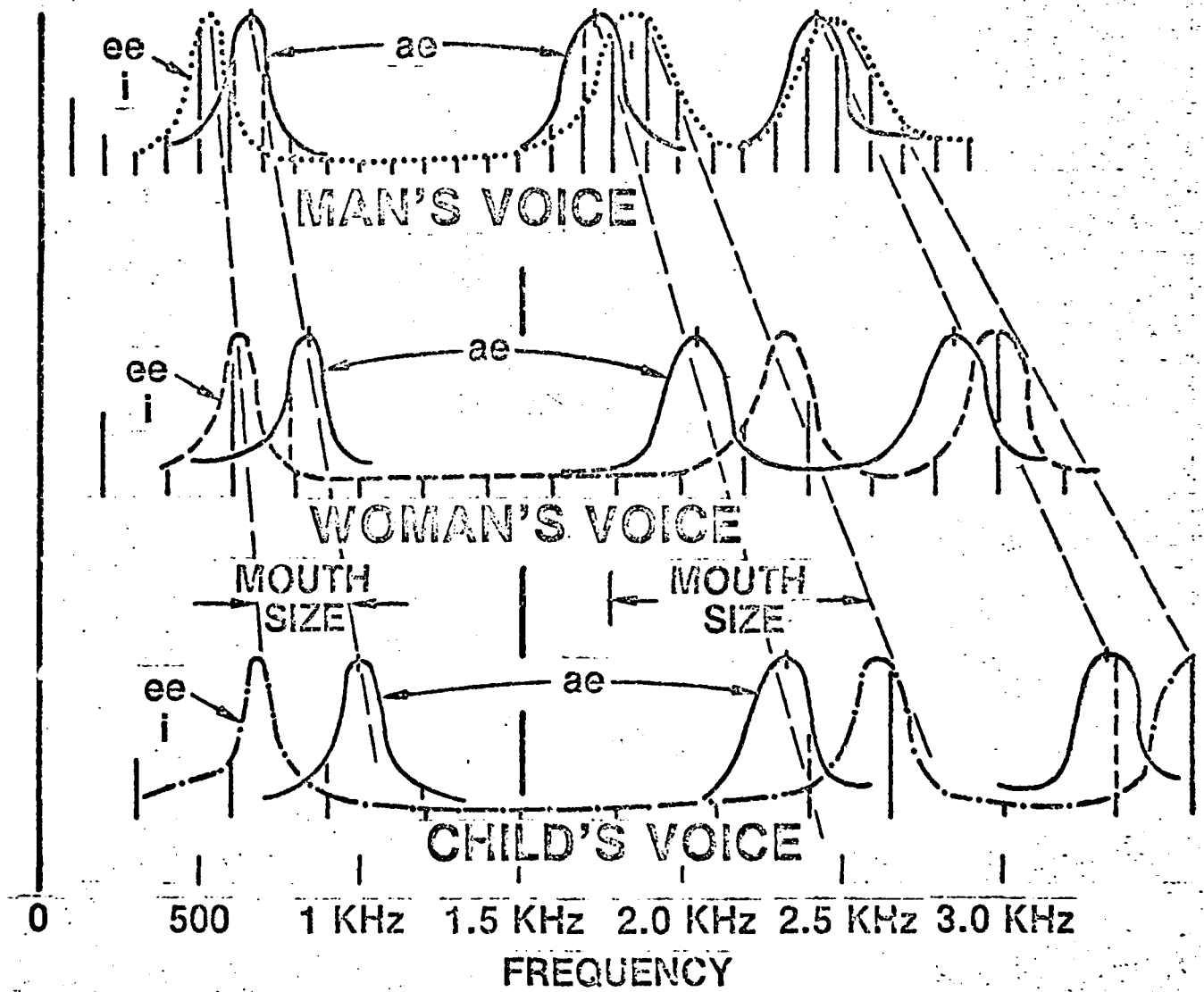
Figure 4.

4

THE PHONEME SPACE
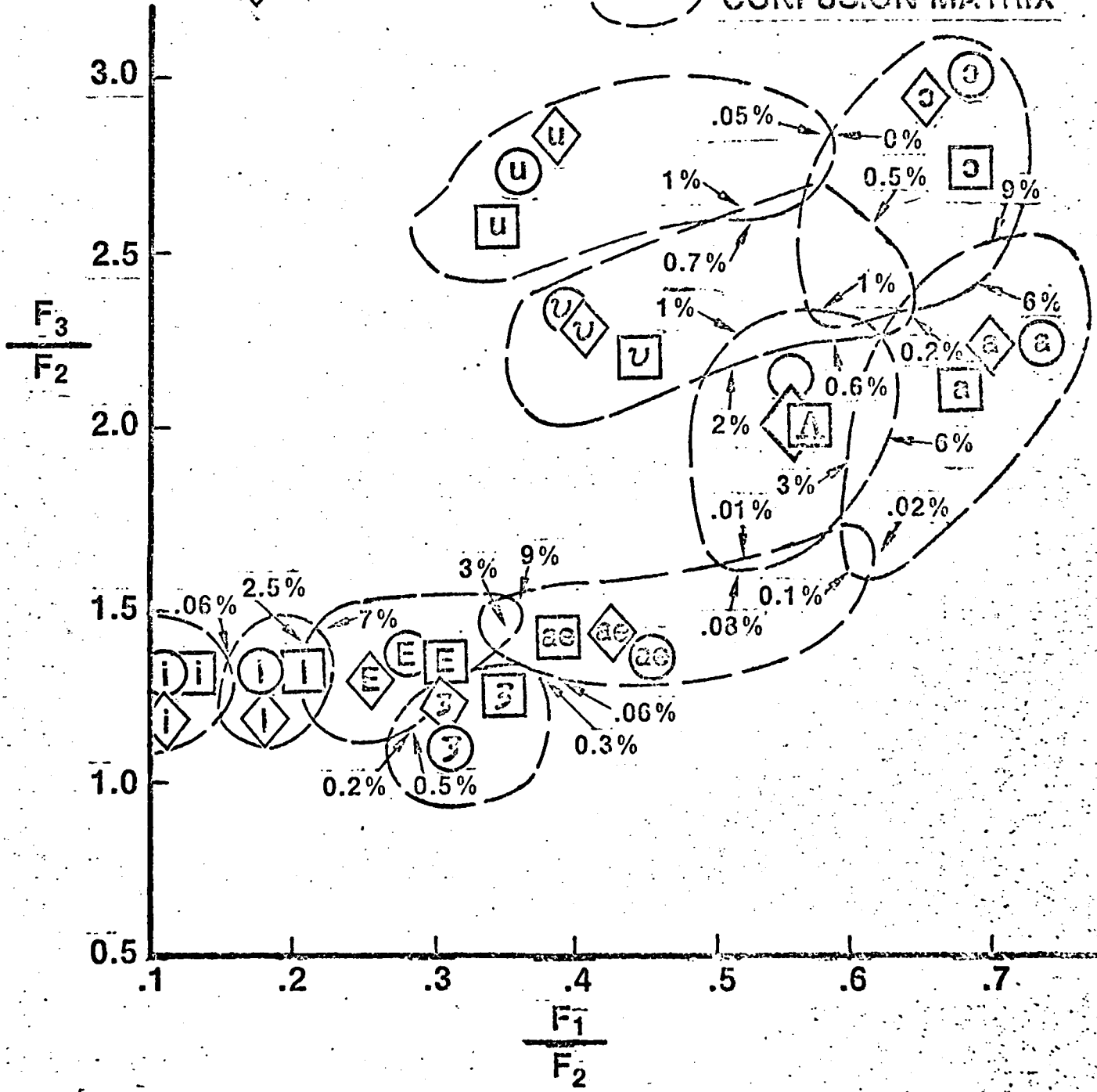
(BASED ON DATA FROM PETERSON AND BARNEY, 1951)

Figure 5   Attachment 13

5

TIME-DOMAIN VOICE DIFFERENCES
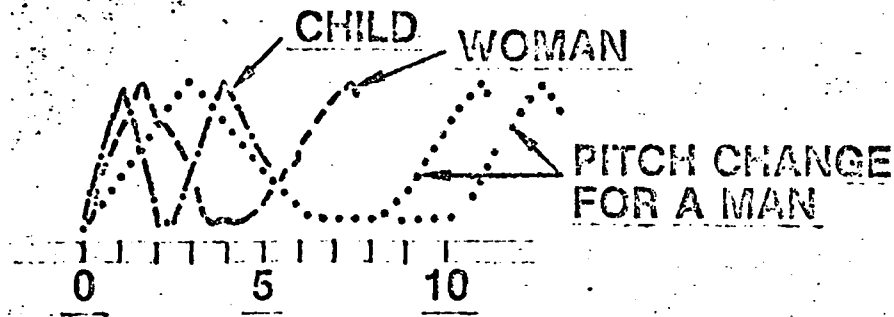
THE GLOTTAL WAVEFORM - MILLISECONDS
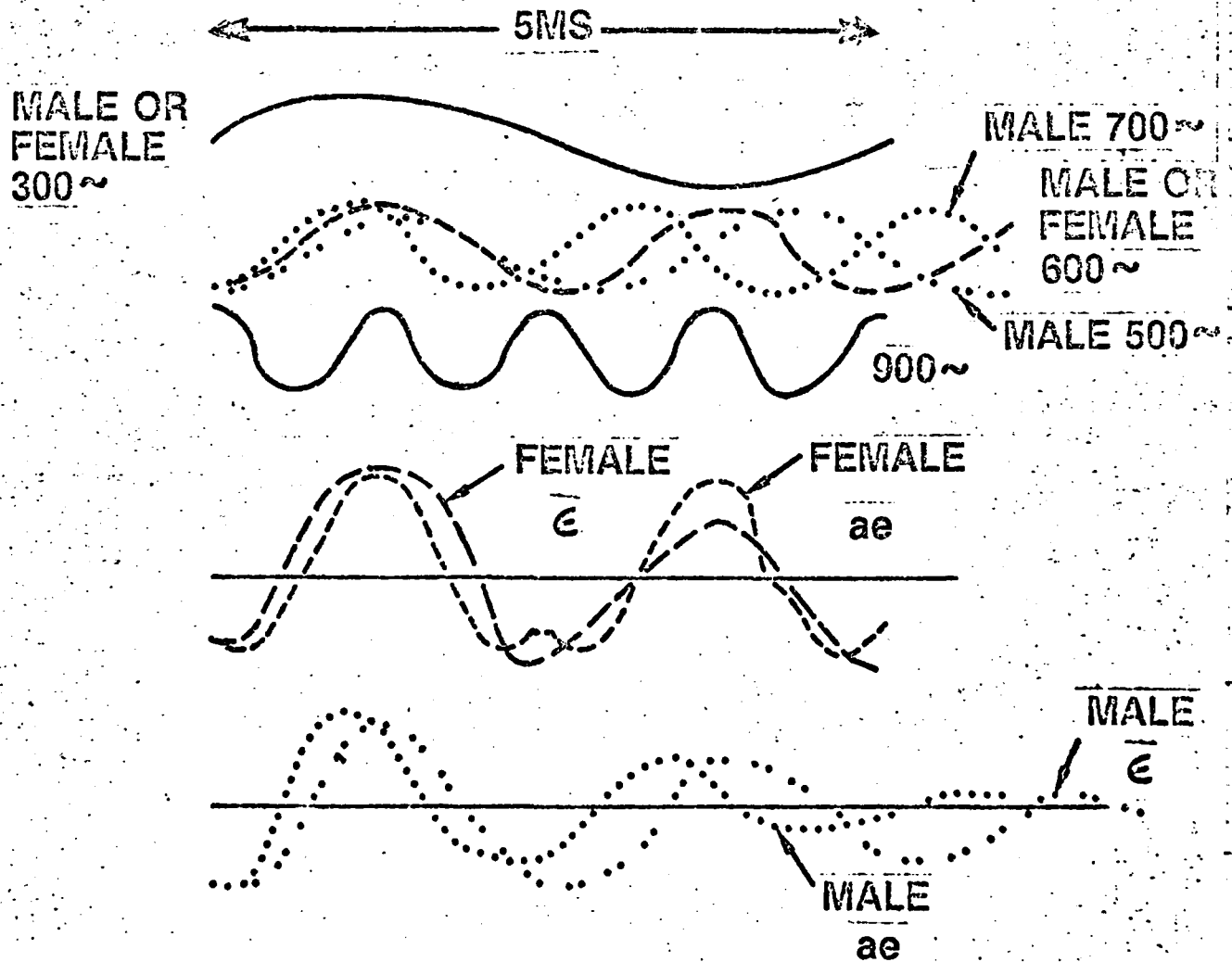
Figure 6

SIMPLIFIED QUANTIZER
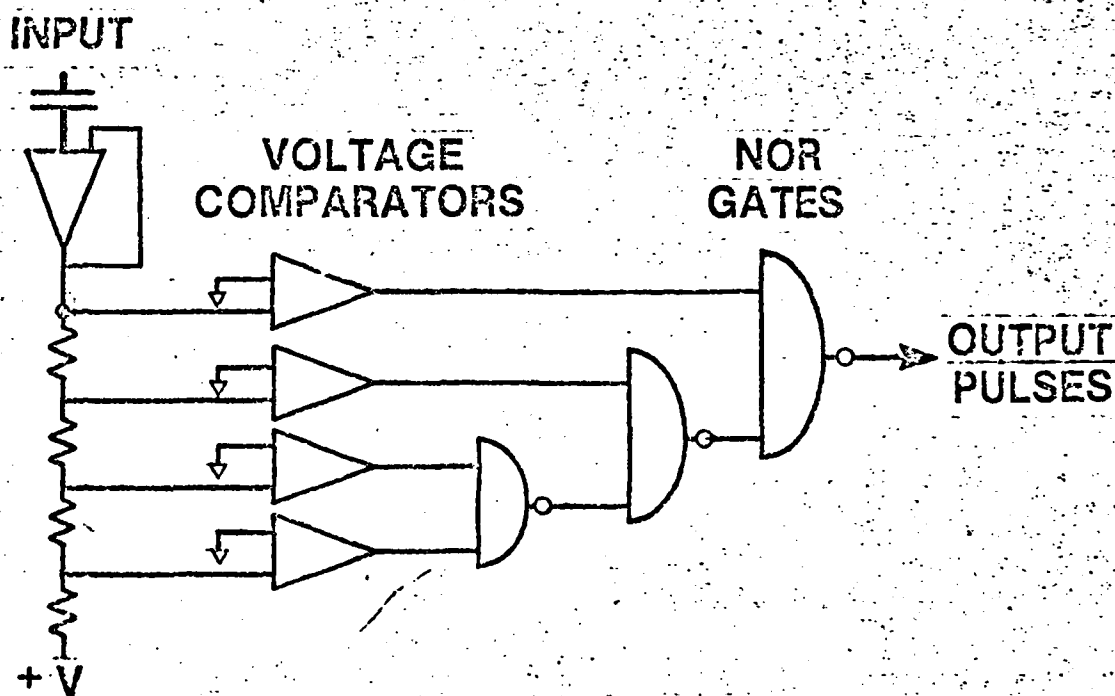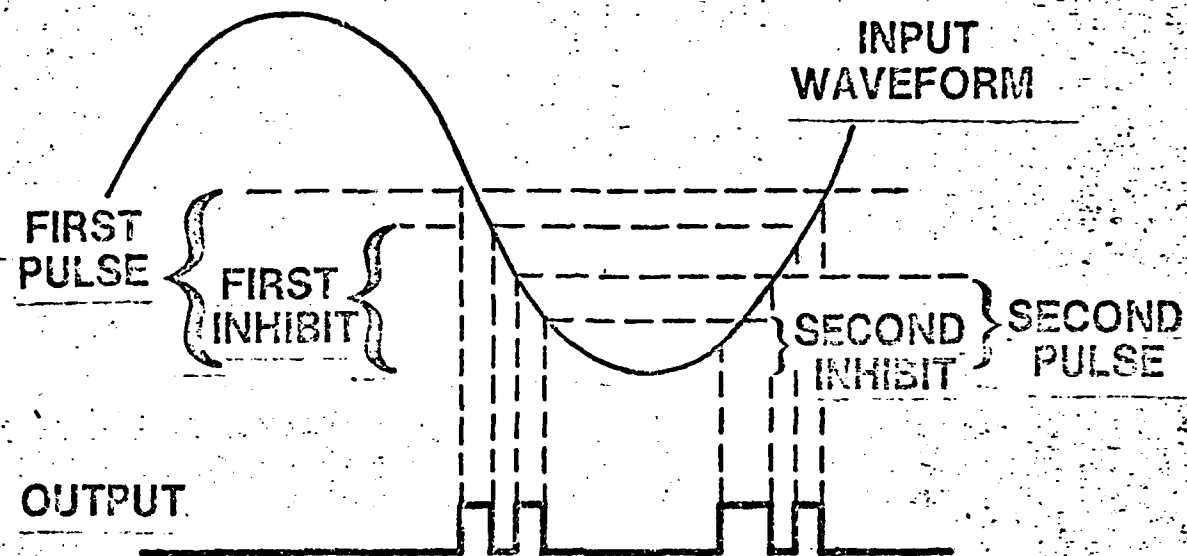
Figure 7.

QUANTIZER TIME DOMAIN
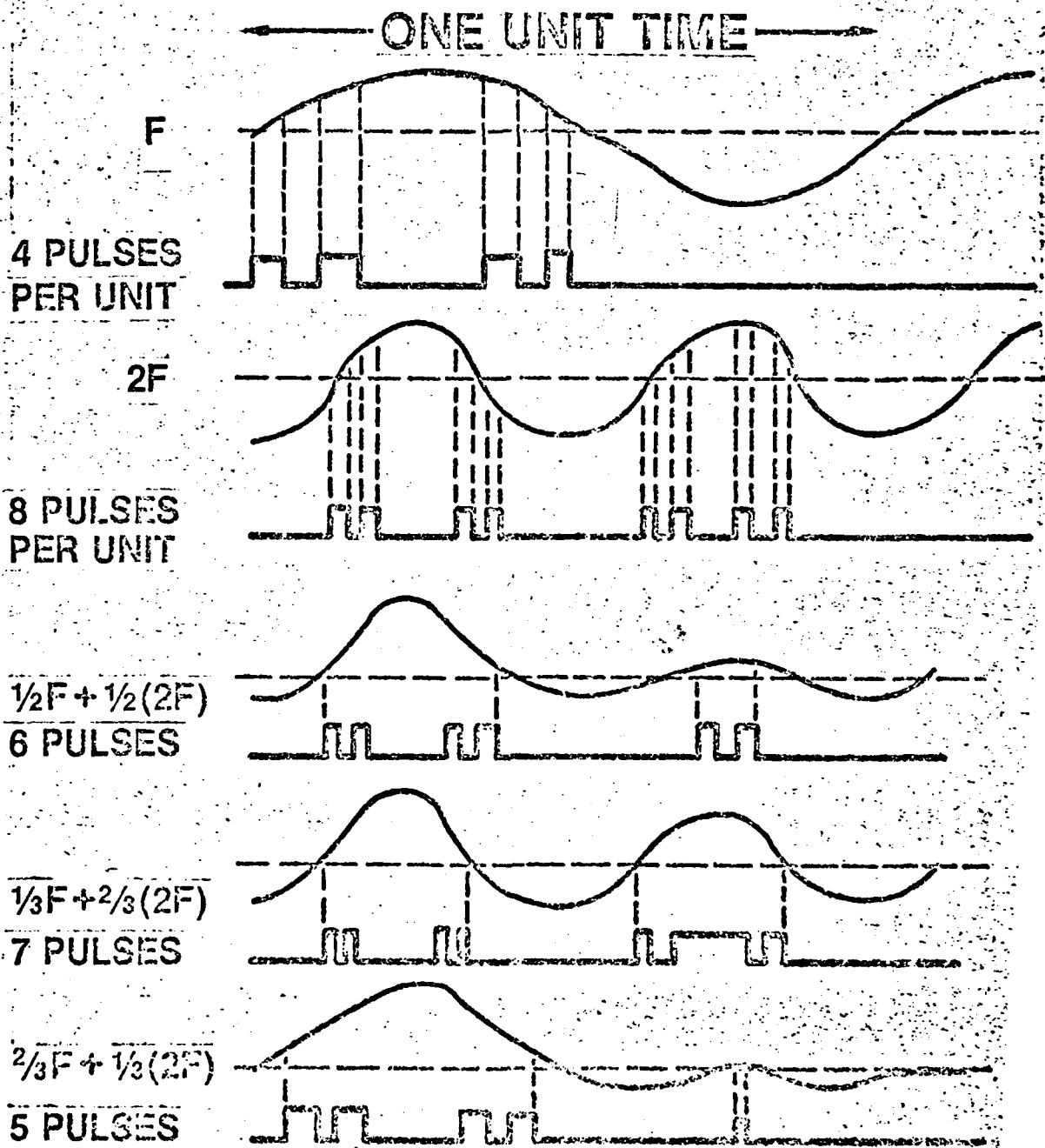
Figure 8

Attachment 13

8

QUANTIZER FREQUENCY DOMAIN
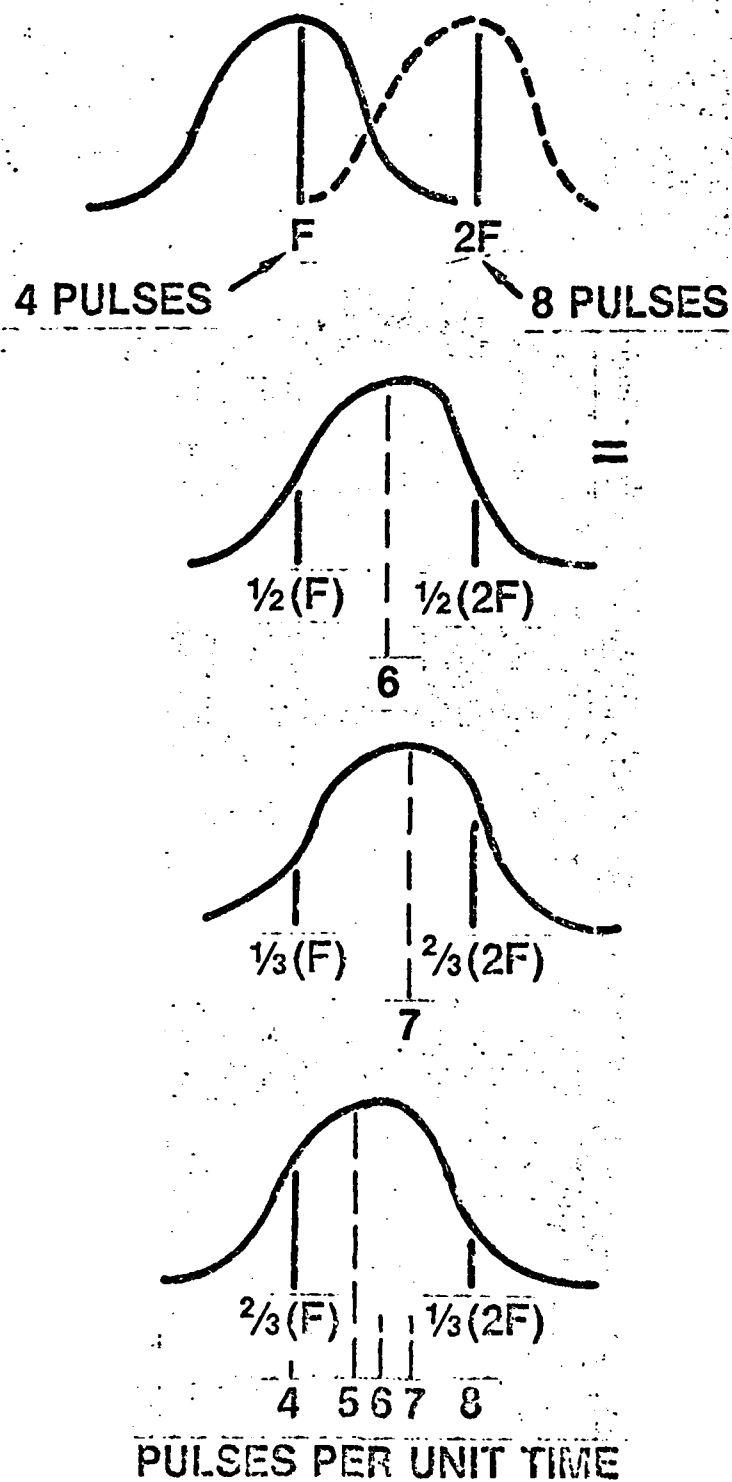
Figure 9.

Attachment 13

VOWEL SYMBOL CHARACTERISTICS

(NOTE THAT ON THE AVERAGE, THE VOWEL RESONANCE BANDWIDTH IS
ABOUT EQUAL TO THE DISTANCE BETWEEN VOWEL FREQUENCIES)
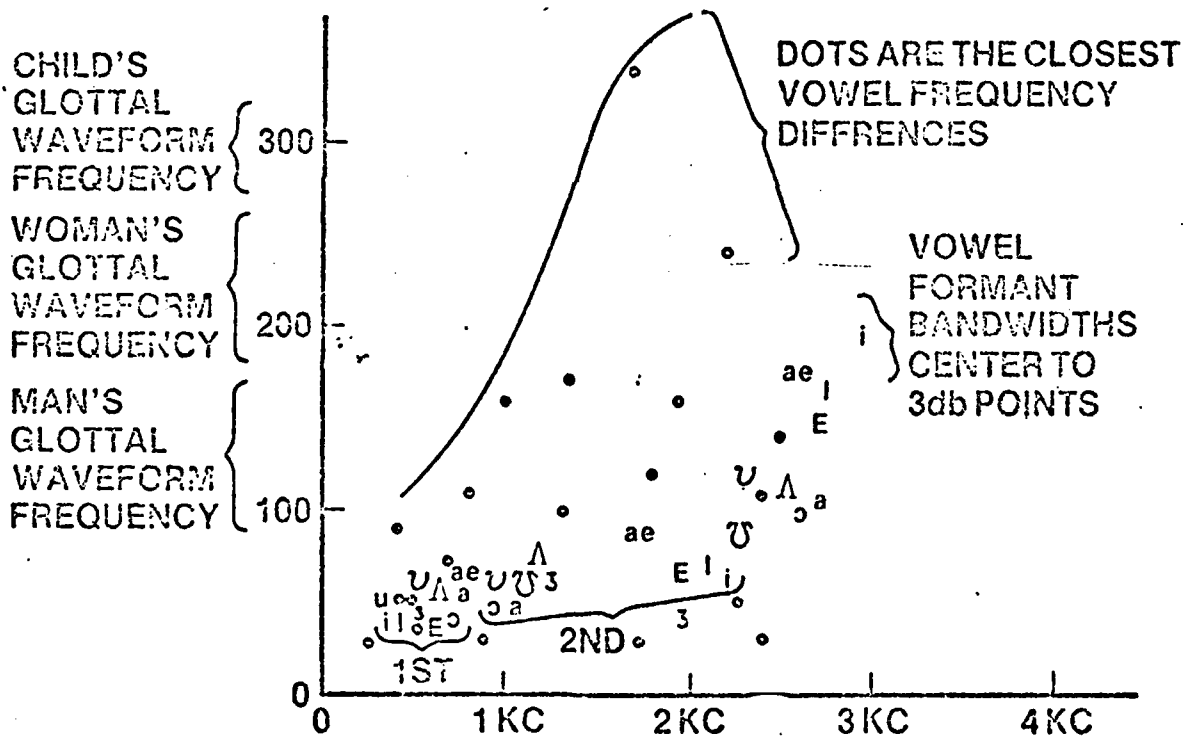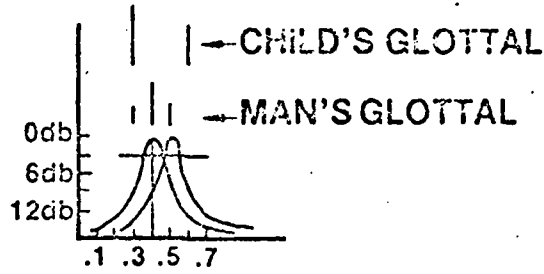
Figure 10

FRICATIVE DETECTION

Figure 11

PLOSIVE DETECTION

Figure 12 (top): PLOSIVE DETECTION — A graph with frequency axis (500 to 3,000) on the vertical and MILLISECONDS (0, 100, 200, 300) on the horizontal. Labels include SILENT INTERVAL, g+K, d+t, b+p, VOICED, and POSSIBLE VOWEL 2ND FORMANTS.

NASAL DETECTION

Figure 12 (bottom): NASAL DETECTION — A graph with frequency axis (500 to 3,000) on the vertical and MILLISECONDS (0, 100, 200, 300) on the horizontal. Labels include ng, n, m, POSSIBLE VOWEL 2ND FORMANTS, and NASAL SOUNDS.
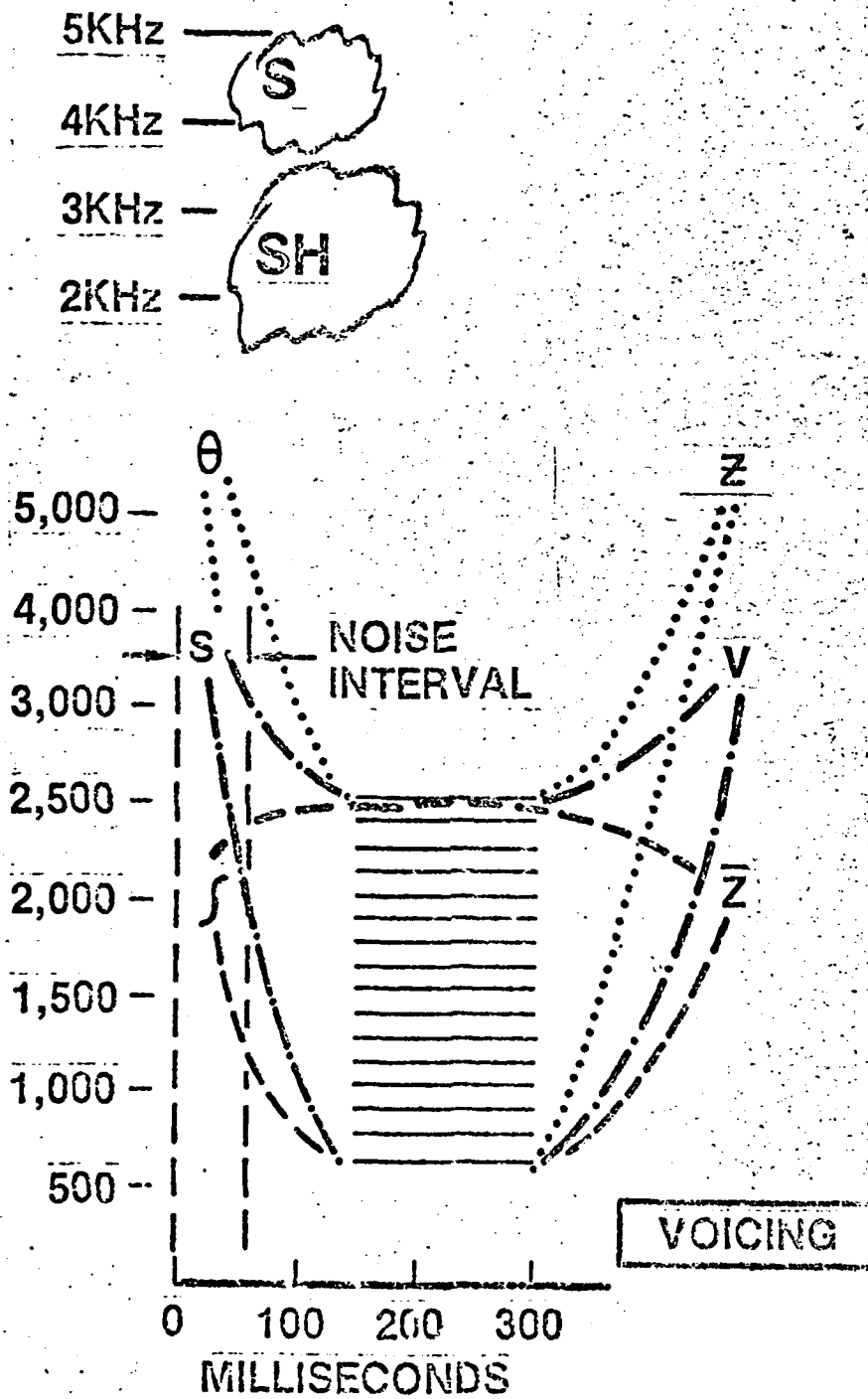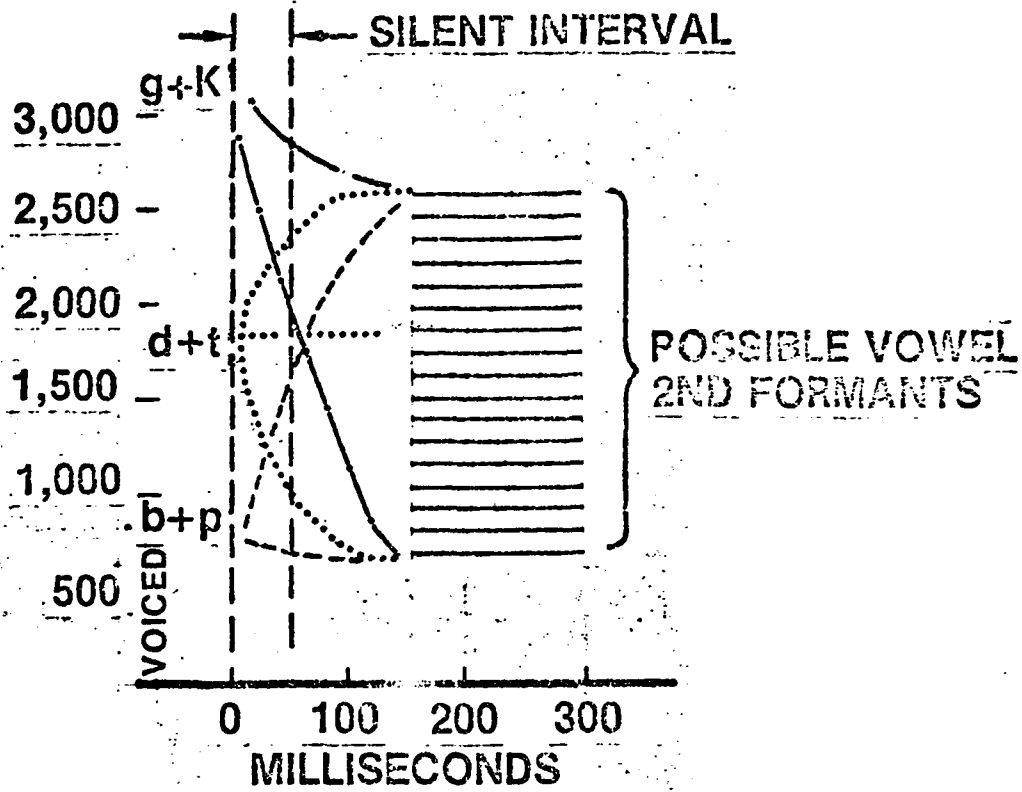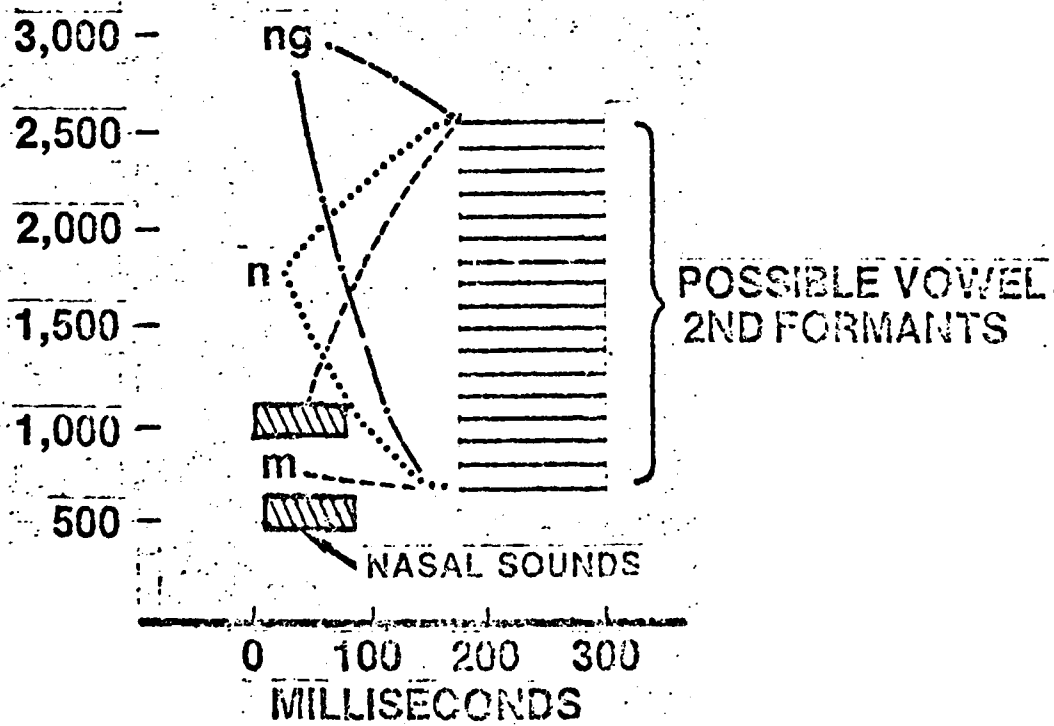
Figure 12.   Attachment 13

12

PHONEME SPACE ESTIMATE
≈ 30 PHONEMES
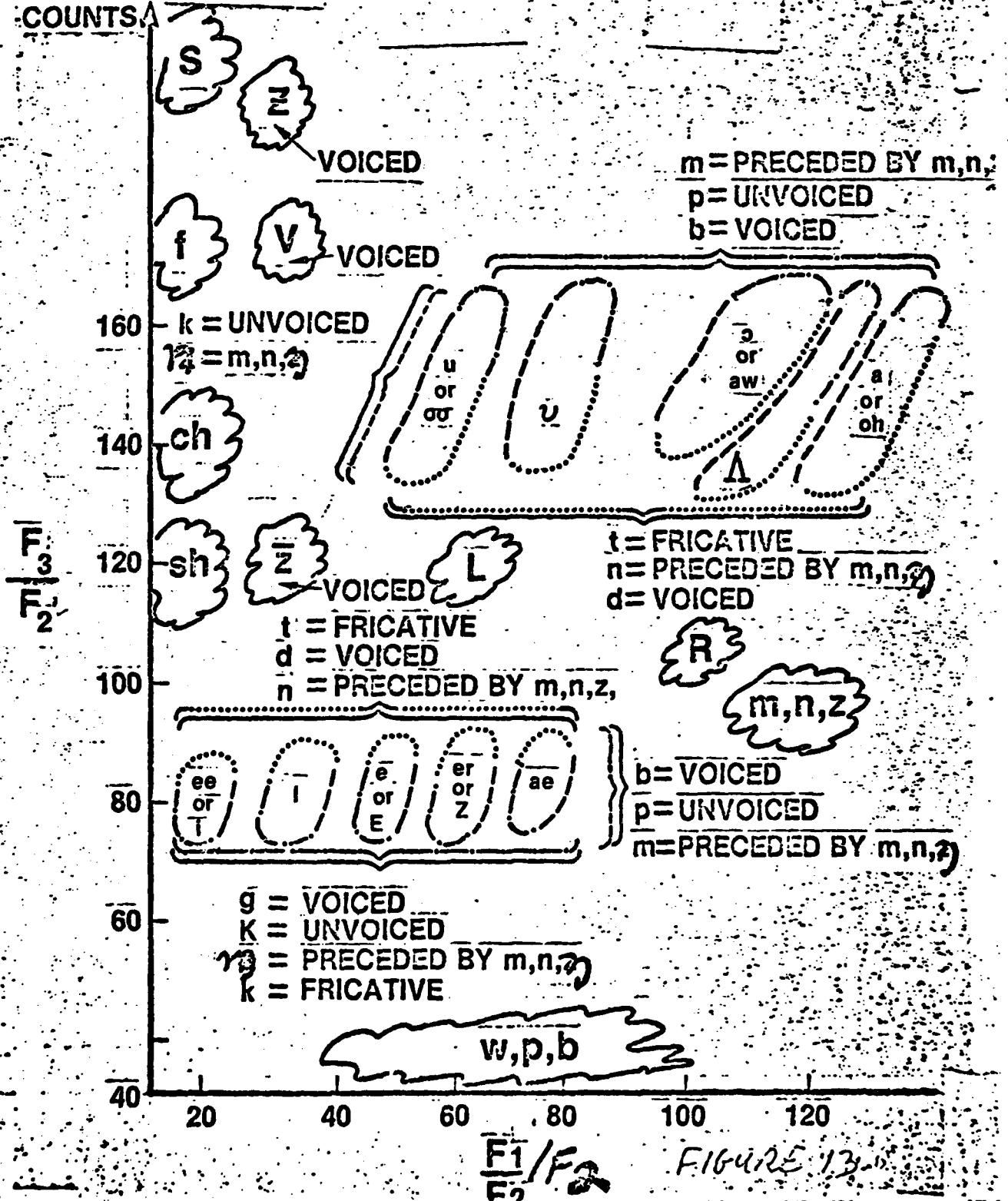
FIGURE 13

BLOCK DIAGRAM OF DECODER

Figure 14

BLOCK DIAGRAM OF DECODER DIGITIZED OUTPUT
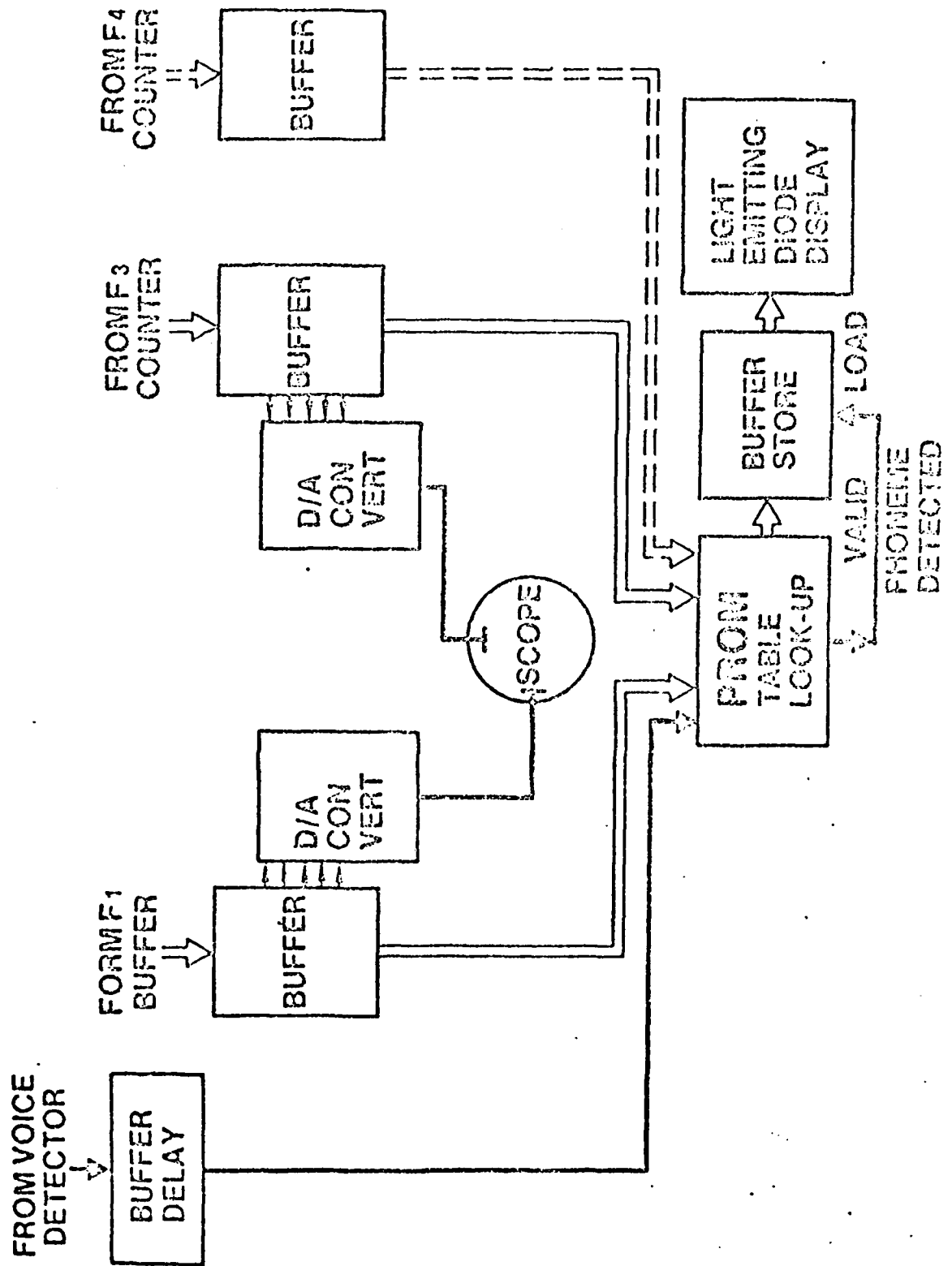
Figure 15

Attachment 13

15

Figure 16

Attachment 13
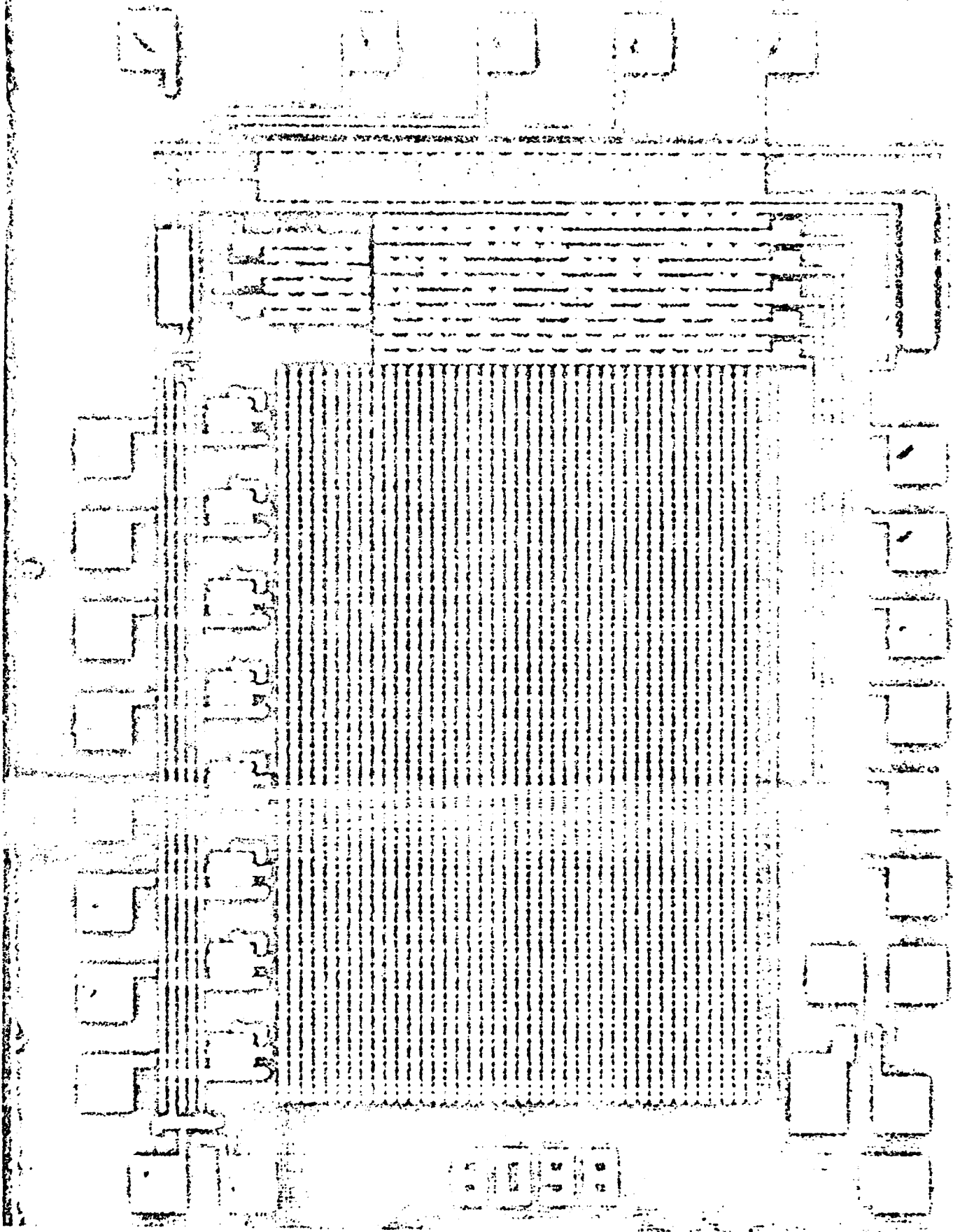
16

TABLE LOOK-UP

16 ROWS

16 OWS

S Z

f V bu b℧ bɔ bΛ ba

u ℧ ɔ Λ a

ch du d℧ dɔ dΛ da

sh Z L

R
m,n,z

d
ae
bae

di bi dı bı de be dʒ bʒ g
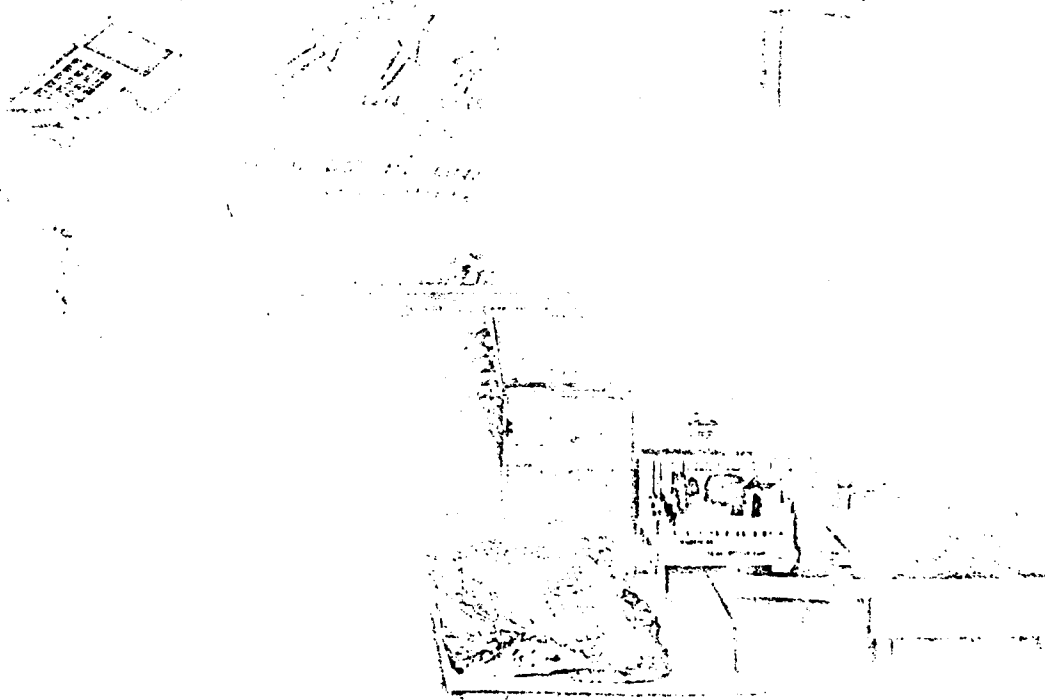gi gı gı gı ge ge gʒ gʒ ae

W B
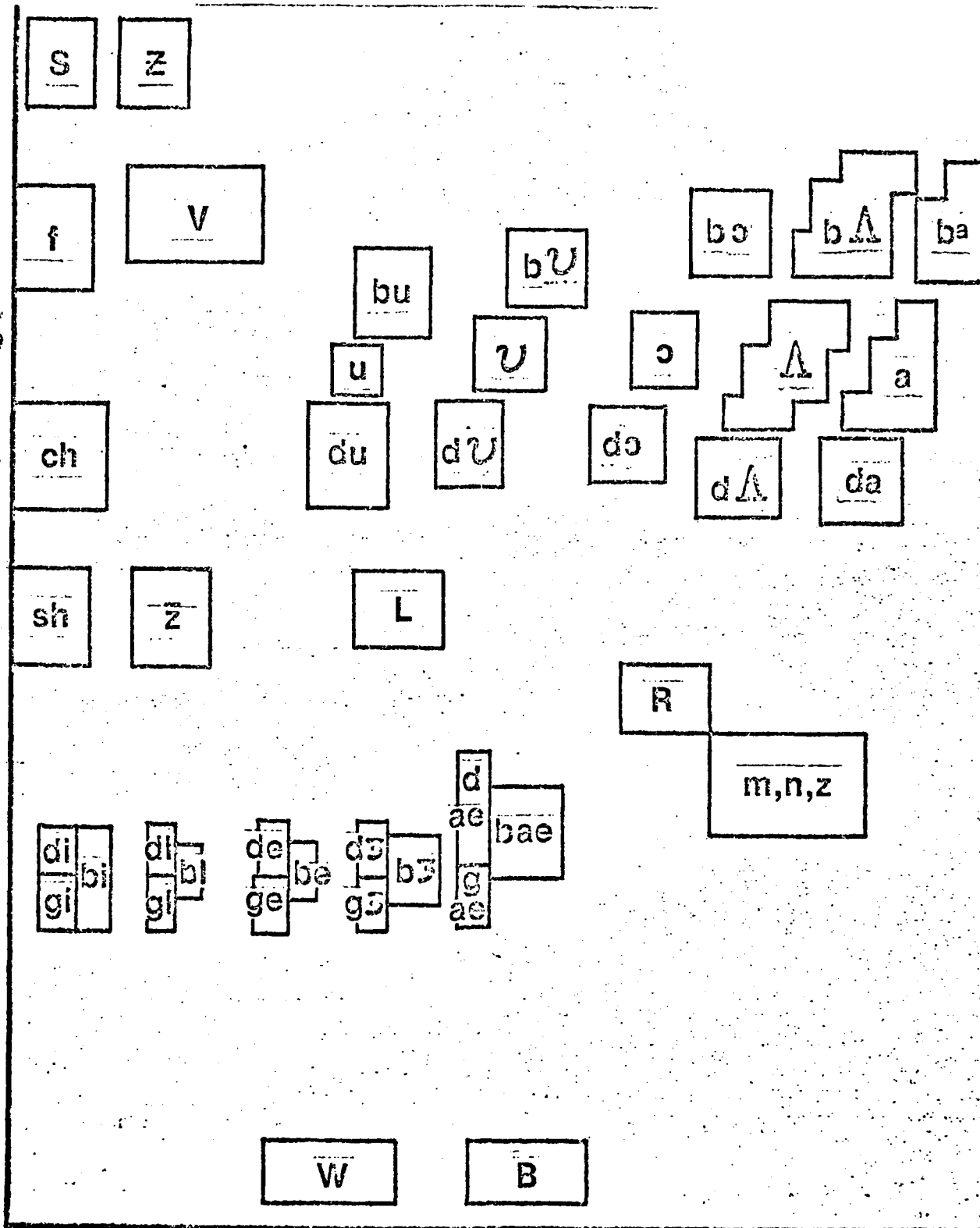
32 COLUMNS

Figure 17

Attachment 13

17

FIGURE 18

A WEARABLE FORM OF THE DEAF COMMUNICATIONS DECODER
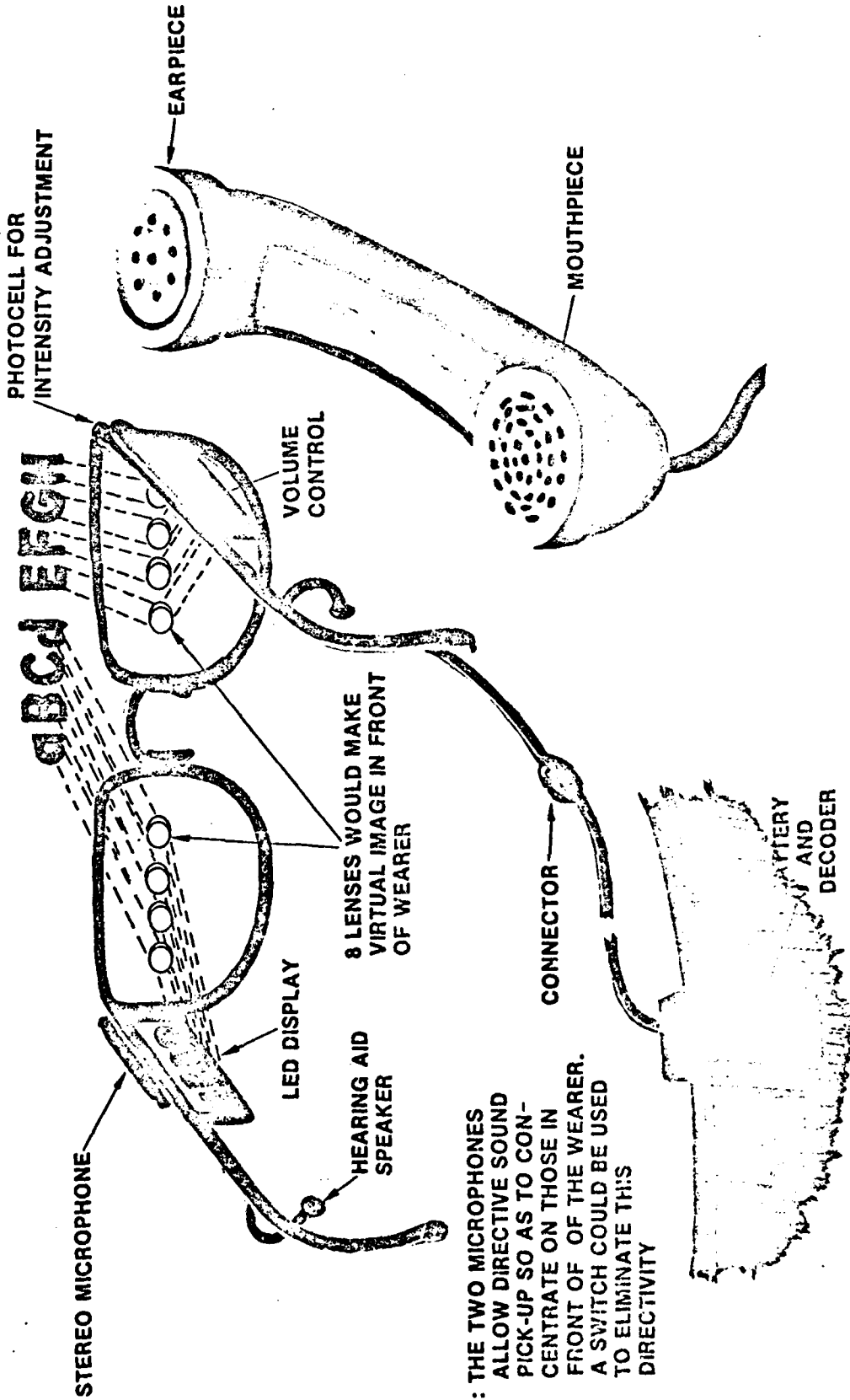
EARPIECE

MOUTHPIECE

PHOTOCELL FOR INTENSITY ADJUSTMENT

VOLUME CONTROL

ABCd EFGH

8 LENSES WOULD MAKE VIRTUAL IMAGE IN FRONT OF WEARER

CONNECTOR

BATTERY AND DECODER

STEREO MICROPHONE

LED DISPLAY

HEARING AID SPEAKER

NOTE: THE TWO MICROPHONES ALLOW DIRECTIVE SOUND PICK-UP SO AS TO CONCENTRATE ON THOSE IN FRONT OF OF THE WEARER. A SWITCH COULD BE USED TO ELIMINATE THIS DIRECTIVITY
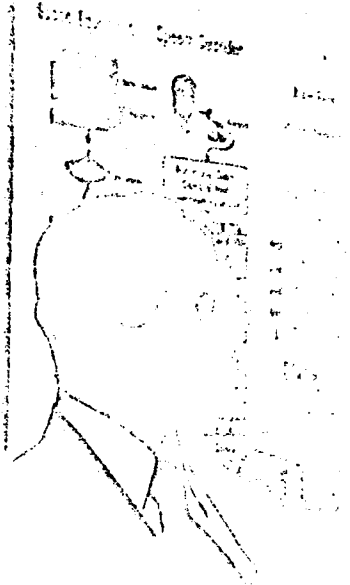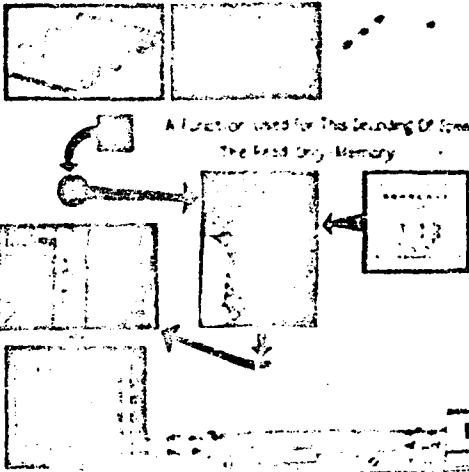
Figure 19. AN EXPANSION OF A CONCEPT BY HUGH UPTON

Attachment 13

19

VOWELS

FRONT   BACK

△ i

y ɨ ω

△ u

↑

MOUTH
OPEN

Ⓘ

ⓤ

MOUTH
CLOSED

∅ ɹ ɣ
æ
ə ʌ

[e]

[o]

ɛ

ɔ

△ a   ʌ

$\square + \triangle$ = ARABIC & ESKIMO

$\square + \triangle$ = SPANISH, GREEK, EGYPTIAN, CZECH, & LATIN

$A + \square + \triangle$ = PERSIAN

$\bigcirc + \square + \triangle$ = ITALIAN & BURMESE ($\triangle + \square$ NASAL)

$\bigcirc + A + \square + \triangle$ = IGBO

$\bigcirc + \bigcirc + \square + \triangle$ = TSWANA

$\partial + \bigcirc + \square + \triangle$ = PORTUGUESE ($\square + i, u, \partial$ NASAL)

$i + \partial + \bigcirc + \square + \triangle$ = N. WELSH

$\jmath + \partial + \bigcirc + \square + \triangle$ = SCOTTISH

$y + \emptyset + æ + A + \bigcirc + \square + \triangle$ = FRENCH ($\square + æ + a$ NASAL)

$i + \omega + \gamma + \Lambda + \bigcirc + \square + \triangle$ = VIETNAMESE

$y + \omega + \emptyset + \square + \triangle$ = TURKISH

$æ + \Lambda + A + \bigcirc + \bigcirc + \square + \triangle$ = ENGLISH
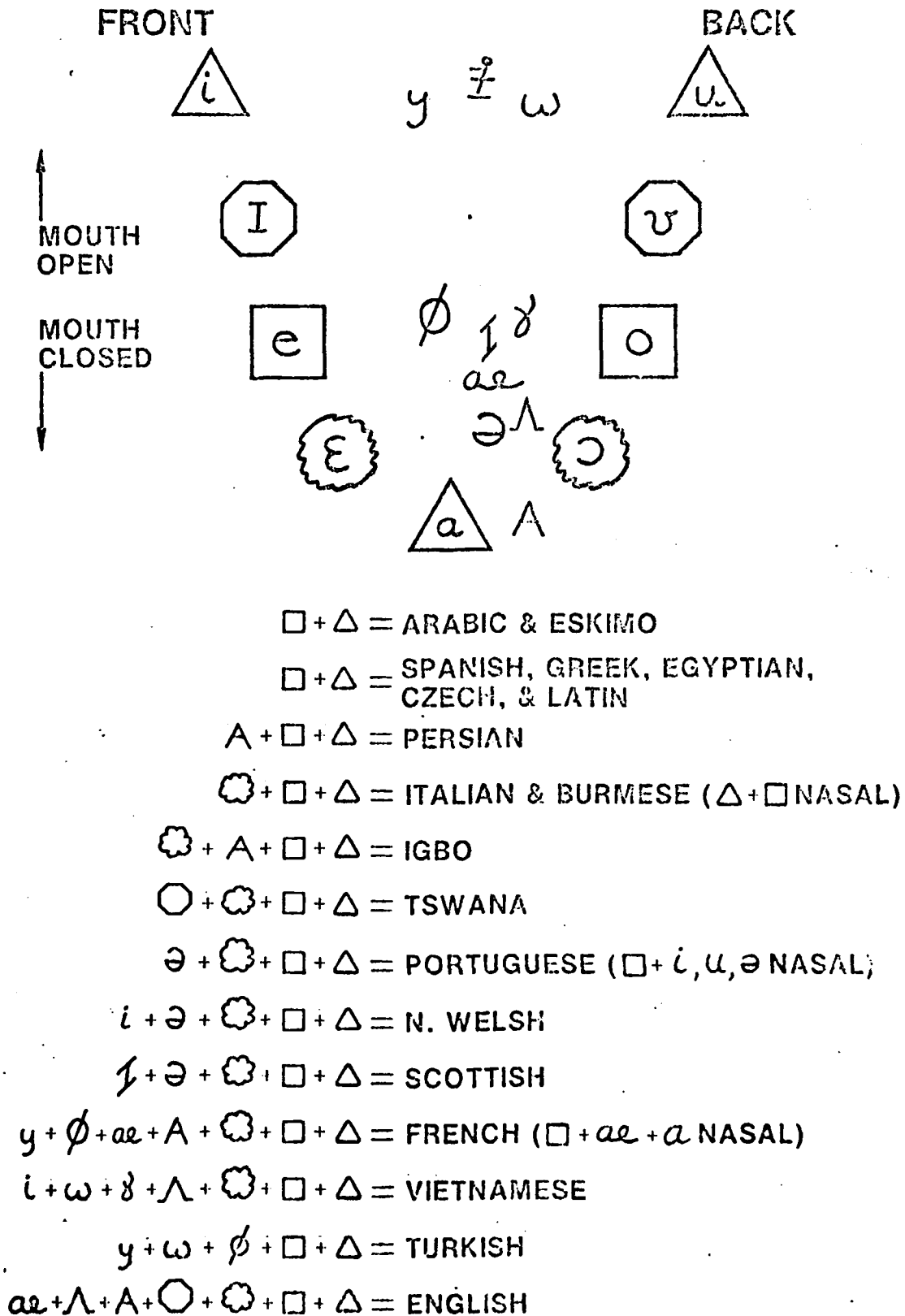
Figure 20

Attachment 13

20

SWEDISH VOWELS (FROM FANT 1959)
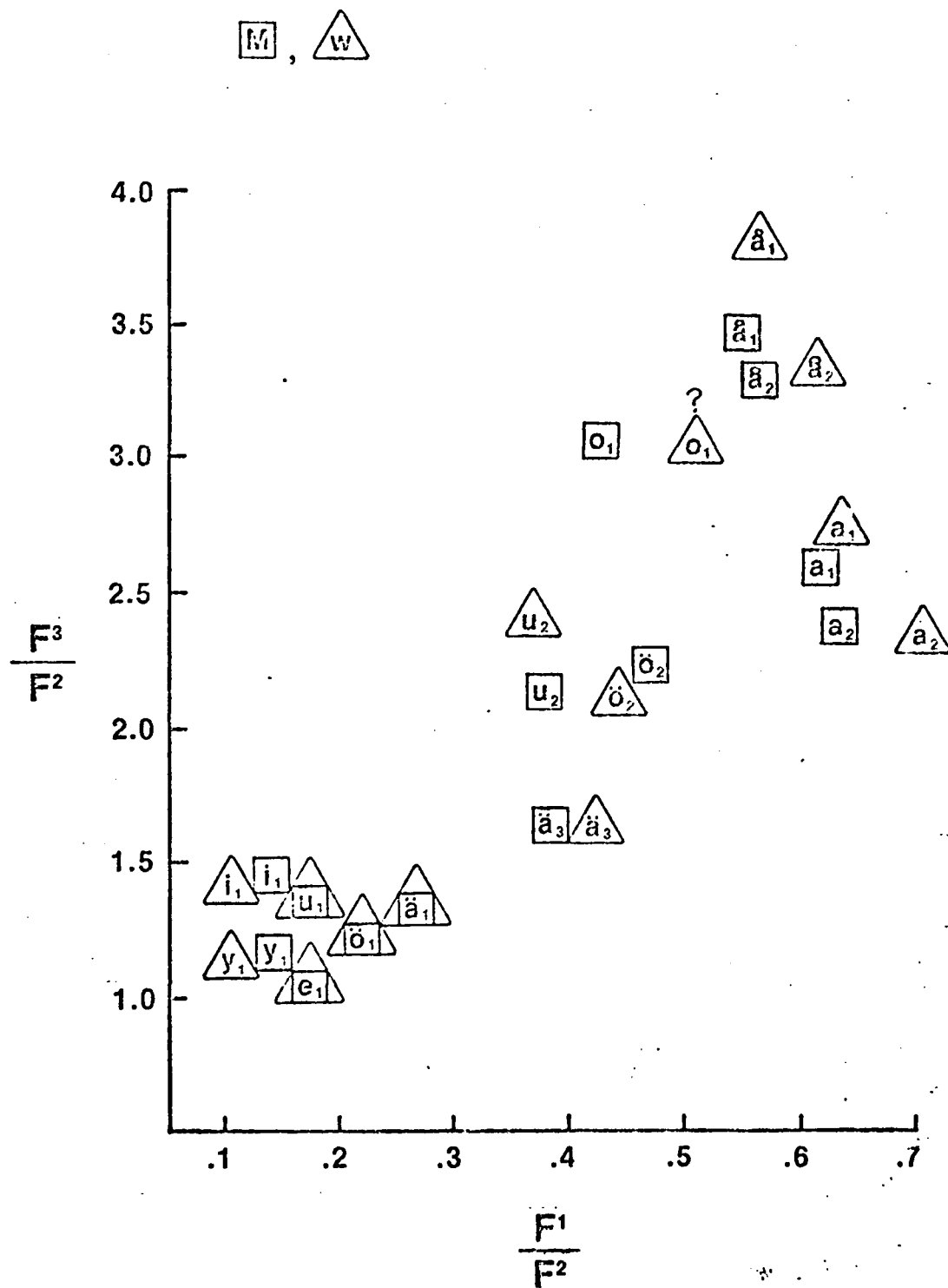
Figure 21

Attachment 13

A phonetic consonant chart with a mid-sagittal diagram of the mouth at top. Column headers (place of articulation): BILABIAL, LABIO-DENTAL, DENTAL, ALVEOLAR, PALATO-ALVEOLAR, RETROFLEX, PALATAL, VELAR, UVULAR, PHARYNGAL, GLOTTAL.

MOST SYMBOLS

STOPS

UNVOICED STRONG LONG FORTIS: P ʔ ƫ t ƫ c k q ʔ

F₂ FOCUS: 800 — 1800 — 3000

VOICED WEAK SHORT LENIS: b ɖ d ɖ ɟ g ɢ

F₂ FOCUS: 800 — 1800 — 3000

NASALS: m n ɲ ŋ ɴ

FRICATIVES

UNVOICED STRONG LONG FORTIS: ɸ f θ s ʃ ʂ ç x χ ħ h

F₂ FOCUS: 3000 5000 2000

VOICED WEAK SHORT LENIS: β v ð z ʒ ʐ ʝ ɣ ʕ ɦ

CONTINUANTS (FRICTIONLESS): l j

Side labels (right):
HUNGARIAN(12)
HINDUSTANI(10)
TAMIL(12)
GEORGIAN(22)
ZULU(38) CLICKS
ZULU(≈19) CLICKS
ESKIMO(4)
ENGLISH(8)
GERMAN(8)

(ADAPTED FROM PHONETICS, BY J. D. O'CONNOR)

Figure 22

Attachment 13

22

WORD RECOGNITION
## AN APPLICATION OF PATTERN MATCHING

TRAINING MODE

WORD 1 → AD CONVERSION & COMPRESSION → DIGITIZED WORD 1 → COMPUTER MEMORY LOC 1 → WORD 1

WORD 2 → AD CONVERSION & COMPRESSION → DIGITIZED WORD 2 → COMPUTER MEMORY LOC 2 → WORD 2

WORD n → AD CONVERSION & COMPRESSION → DIGITIZED WORD n → COMPUTER MEMORY LOC N → WORD n

WRS MEMORY

WORD 1
WORD 2
--
WORD n

LIVE MODE

WORD i → AD CONVERSION & COMPRESSION → DIGITIZED WORD i → WORD i

TABLE LOOK UP FOR MOST LIKELY MATCH

WORD 1
WORD 2
WORD i
WORD n

UNITED STATES ARMY SYSTEMS TACTICAL DATA

Attachment 14

TYPICAL TACFIRE MESSAGES

MESSAGE	1st FIELD	FIELDS	NUMERIC
FIRE MISSION GRID	EASTING	15	9
FIRE MISSION SHIFT	FROM	18	13
FIRE MISSION POLAR	DIRECTION	14	9
ADJUST	TARGET	12	7
SHELL REPORT	DIRECTION	8	6
FIRE MISSION QUICK	POINT	4	1
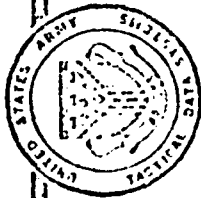
TRAINING CONSIDERATIONS

* SELECTABLE, RESTRICTED POPULATION
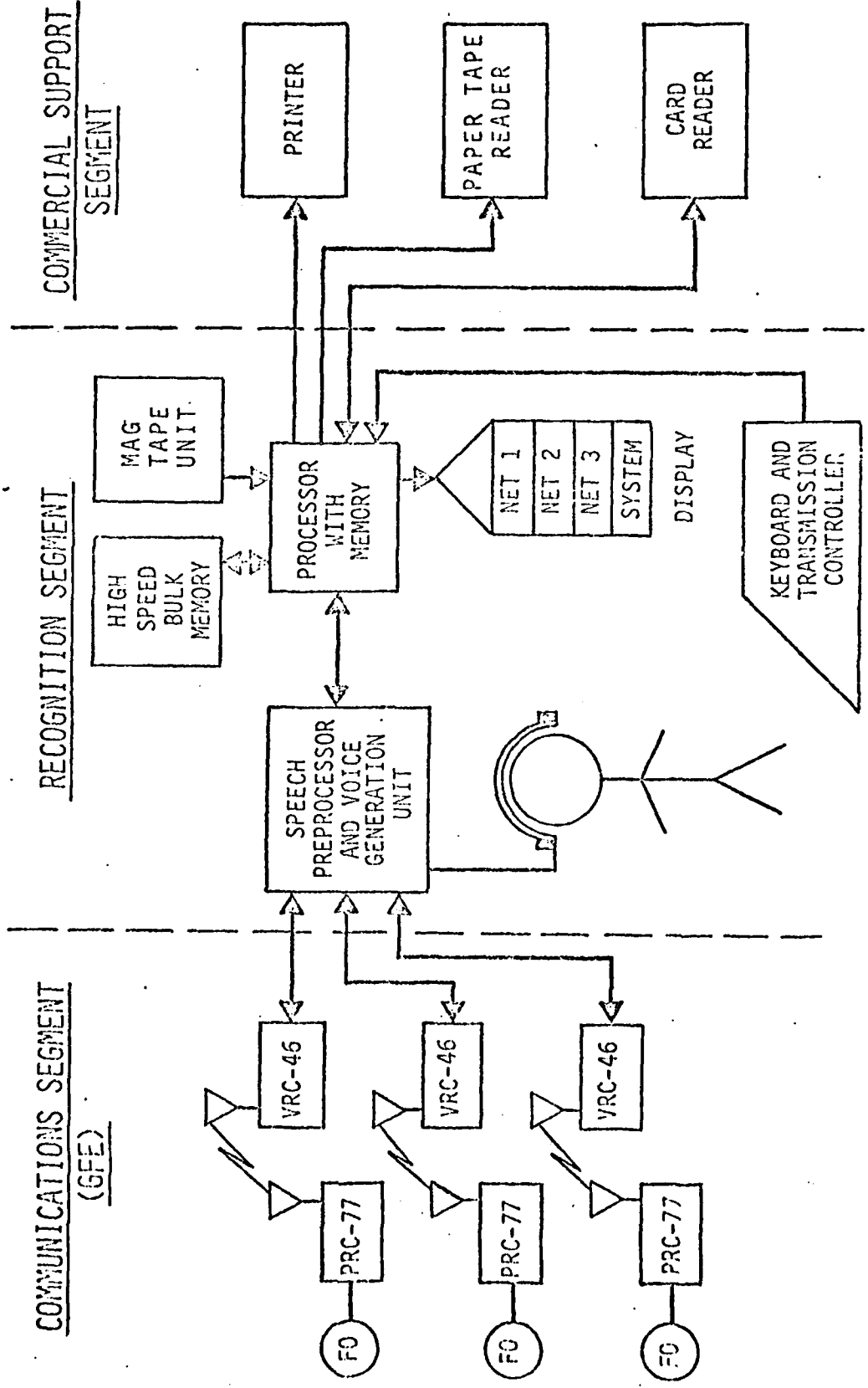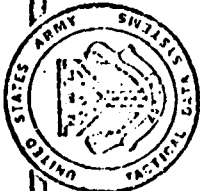
* MODIFIABLE VOCABULARY

* PREPONDERANCE OF DIGITS

REQUIREMENTS VERSUS STATE OF THE ART

- ○ ACTIVE VOCABULARY OF 36 WORDS: WELL DEMONSTRATED

- ○ ACCURACY OF 95% AT 10db S/N (15db = POOR)

  - ○ WELL DEMONSTRATED WITH HIGH QUALITY INPUT

  - ○ OPERATION OVER TACTICAL FM NETS: ECOM DEMONSTRATION

  - ○ PRE-AWARD REQUIREMENT: 90% AT 15db S/N

- ○ VERIFY SPEAKER: DESIGN GOAL OF 2% ERROR; SECONDARY CHARACTERISTIC

- ○ PROMPTING: NUMEROUS EXISTING SPEECH SYNTHESIS DEVICES

- ○ EXISTING MILITARIZED EQUIPMENT: A NUMBER OF SOURCES

TACTICAL WORD RECOGNITION SYSTEM

COMMERCIAL SUPPORT SEGMENT

- PRINTER
- PAPER TAPE READER
- CARD READER

RECOGNITION SEGMENT

- MAG TAPE UNIT
- HIGH SPEED BULK MEMORY
- PROCESSOR WITH MEMORY
- NET 1 / NET 2 / NET 3 / SYSTEM DISPLAY
- KEYBOARD AND TRANSMISSION CONTROLLER
- SPEECH PREPROCESSOR AND VOICE GENERATION UNIT

COMMUNICATIONS SEGMENT (GFE)

- VRC-46
- VRC-46
- VRC-46
- PRC-77
- PRC-77
- PRC-77
- FO
- FO
- FO

MESSAGE ENTRY CONCEPT (OVER TACTICAL RADIO)

USER		WRS	
A C THIS IS B1	OVER	B1 THIS IS A C	OVER
FIRE MISSION GRID	OVER	FIRE MISSION GRID NORTHING	OVER
1 2 3 4	OVER	1 2 3 4 EASTING	OVER
9 8 7 6	OVER	9 8 7 5 SHELL	OVER
CORRECTION 9 8 7 6	OVER	9 8 7 6 SHELL	OVER
HIGH EXPLOSIVE	OVER	HIGH EXPLOSIVE FUZE	OVER
DELAYED	OVER	DELAYED	OUT

WORD RECOGNITION
LIMITED VOCABULARY, LIMITED SPEAKER
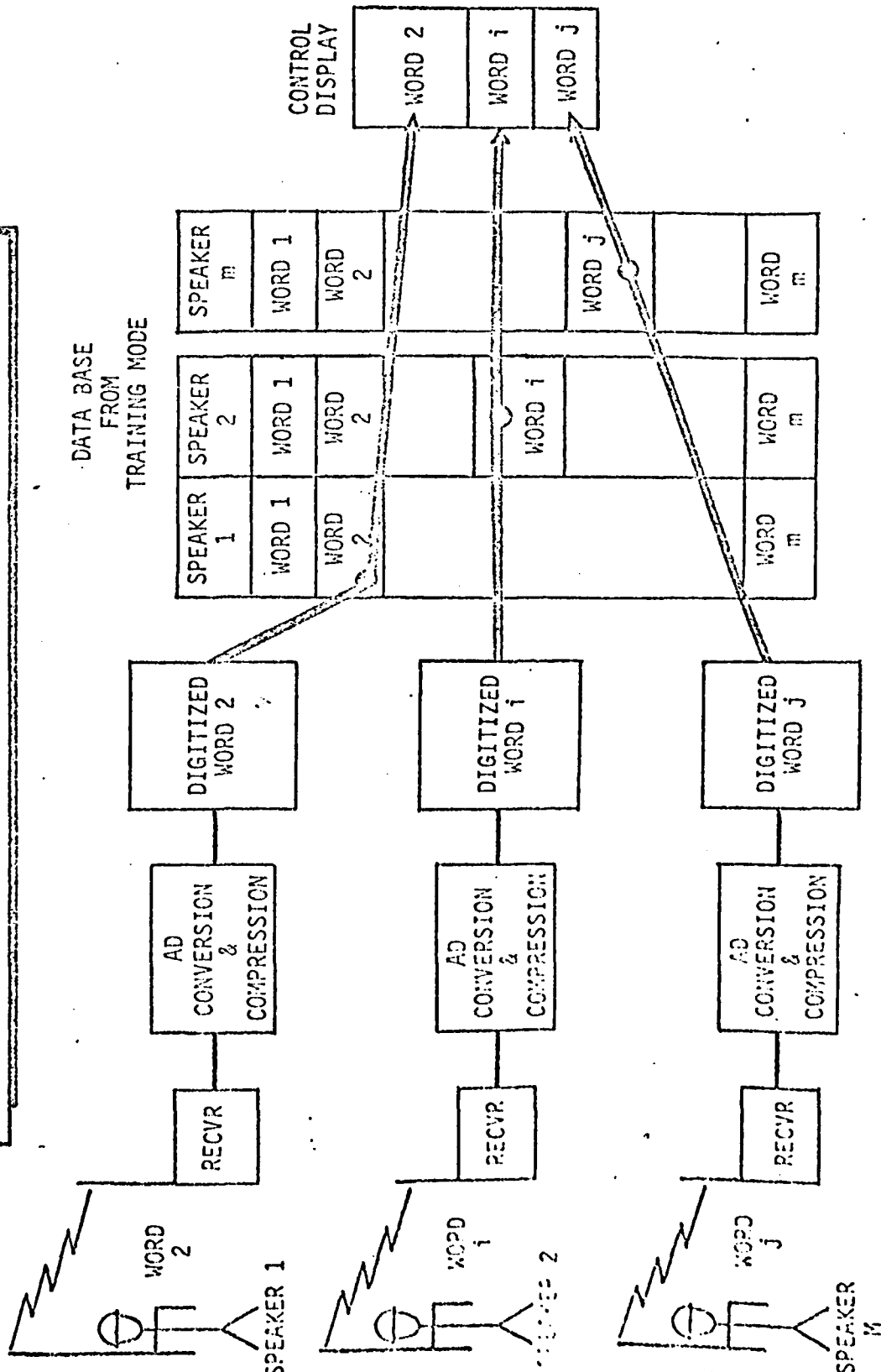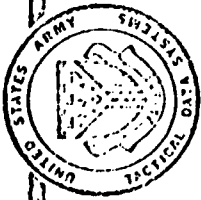
WRS TEST ELEMENTS

* ACCURACY AND RATE OF DATA ENTRY

* EFFECT OF FATIGUE, STRESS, AND ELAPSED TIME

* EFFECT OF RADIO VARIATIONS AND SECURE EQUIPMENT

* SYSTEM INITIALIZATION TECHNIQUES AND TRAINING

END

FILMED

2-83

DTIC