BOLT BERANEK AND NEWMAN INC

CONSULTING • DEVELOPMENT • RESEARCH

BBN Report No. 3093 June 1975

COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

Part I. Packet Radio
II. Speech Compression
III. Vocoded-Speech Evaluation

Quarterly Progress Report No. 2
1 March 1975 to 31 May 1975

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 2935. Contract No. MDA903-75-C-0180.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: BBN Report No. 3093
2. GOVT ACCESSION NO.: AD-A155 236
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY
5. TYPE OF REPORT & PERIOD COVERED: Quarterly Progress Report 1March 1975-31 May 1975
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Jerry Burchfiel, R. Viswanathan, Raymond S. Nickerson
8. CONTRACT OR GRANT NUMBER(s): MDA903-75-C-0180
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Bolt Beranek and Newman Inc. 50 Moulton St. Cambridge, Mass. 02139
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS:
12. REPORT DATE: July 1975
13. NUMBER OF PAGES: 75
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office):
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES: This research was supported by the Advance Research Projects Agency under ARPA Order No. 2935.
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Packet radio; computer communications; packet switched networks; cross-network debugging; ELF; BCPL; speech compression; vocoder; linear prediction; parameter quantization; Optimal parameter interpolation; Gateway; Real time signal processing; Multidimensional scaling
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

This document describes progress on (1) the development of a packet radio network, (2) speech compression and (3) speech evaluation. Activities reported under (1) include work on ELF, VM ELF, the IMP-11A interface, cross-network debugger and the PDP-11 gateway; under (2) investifation of the issues involving linear predictor gain parameters, and testing of a 1000 bits/second system; and under (3) application of multidimensional scaling techniques to quality evaluation.

Additional keywords: computer communications; packet switched networks; vocoders; parameter quantization; real time signal processing.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

BBN Report No. 3093 June 1975

COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

Part I. Packet Radio

Quarterly Progress Report No. 2
1 March 1975 to 31 May 1975

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 2935. Contract No. MDA903-75-C-0180.

TABLE OF CONTENTS

I. Packet Radio Activities

A. Documentation

During this quarter we released three documents describing our packet radio activities:

1. Packet Radio Temporary Note (PRTN) #138, "Packet Radio Station Hardware, Operating System and Applications Programming Environment" This memo describes the hardware configuration and the software environment which support the programs to implement station functions.
2. PRTN #141, "X-NET Cross-network Debugger User's Manual" Since X-NET employs new techniques in debugging, this manual puts special emphasis on network interaction and other aspects peculiar to the X-NET debugger.
3. PRTN #142 (also RFC 685), "Response Time in Cross-network Debugging" Response time due to delay through the ARPA Network is found not to preclude this type of debugging.

B. Meetings

We attended a March 6-7 meeting at ARPA to define and describe the various protocols to be used in the Packet Radio Network. All contractors were in attendance, and consensus was reached on the number of levels of protocols to be used, their names, and an initial description of the characteristics of each.

At the next working group meeting, May 21 in Anaheim, each contractor gave a detailed progress report with future plans, and

the System Engineering and Technical Direction Group presented detailed plans and schedules for work through the end of the calendar year.

C. Modifications to ELF -- PRI, PRO, XNCP, debugger mods

The latest ELF release (version 75C.1) has been put up on our PDP-11 and is in use. We run the system configured without virtual memory. It is the standard ELF kernel except for two modifications we have installed. First, a small change (resetting the breakpoint trap vector during system initialization) was apparently overlooked by SCRL when they installed other BBN changes; we placed the change in our current ELF. Second, a problem due to the stack pointer being reset when breakpoints are hit was solved by making the system breakpoint handler decrement the program counter (if the trap is not a trace trap), rather than decrementing it in the debug process. This latter modification we view as a temporary measure. The ultimate solution will probably involve more extensive modification, including an "accumulator block" concept similar to that in other operating systems. (Currently ELF makes no distinction between the registers of a process running its own instructions and running instructions of a system call. We defer further comment since this topic is subsumed by the ELF evaluation mentioned below.)

SCRL's ELF NCP and XNCP have been completely replaced by our own XNCP (experimental network control program). Ours occupies only a little over 500 words of code, plus buffers. Currently it is configured to maintain a maximum of two open connections. Since it

is not burdened with a complex of resets during initialization, it is ready for network traffic almost immediately after ELF is started. This speed and its reliability make our XNCP far more usable than the SCRL version. Besides removing the old NCP and XNCP and installing our XNCP, a modification was made to the network interface driver (KDIMP module of ELF) to run the ANTS interface in the same byte-swapping mode as the IMP11A. This allowed a simplification to the cross-network debugger's server process, which has been rewritten to take advantage of the change.

ELF device driver code for the IMP11A interface has been written, debugged and installed in the ELF we run. The IMP11A appears as devices PRI (Packet Radio Input) and PRO (Packet Radio Output). The driver attempts recovery from "ready line error" conditions both by retransmitting outbound messages and by discarding inbound messages. This is to be compared with SCRL's ANTS interface driver, which merely reports an error whenever such a condition arises. The BBN IMP11A driver retransmits five times before reporting a transmit error condition. Our driver has been released to the ELF community with installation instructions and suggestions on modifying the SCRL NCP and XNCP to use the IMP11A.

D. Progress with VM ELF

We are starting an evaluation of virtual memory ELF. We will build a virtual memory ELF for our PDP-11 configuration and verify that it runs. We will then examine issues including: installing our XNCP in the VM ELF; starting the XNCP in its own address space; and

making the cross-network debugger process function in a virtual memory environment. This latter may include assessing the "accumulator block" problems mentioned earlier.

E. IMP-11A Interface

The IMP11A hardware has been debugged. The changes were communicated to the manufacturer (DEC), which has resulted in ECOs issued by DEC for these changes. The problems were traced to deviations from the original design, introduced by DEC in the process of repackaging. The deviations resulted in noise in the handshake with the IMP. Also, some open lines to the DR-11B were causing extraneous non-processor request direct memory accesses.

We have two versions of an IMP11A test program: one runs stand-alone, and one runs under ELF. Using hardware loop-back in the IMP11A, the standalone diagnostic has run continuously for 17 hours, during which it transferred over a billion 8-bit bytes, at an average rate of 147000 bits per second, with no errors.

With this result, we are now ready to literally plug in the digital PR unit from Collins and run the test on it.

F. Specification for Basic Station modules

We have completed the general layout of software modules in the station. This layout, and preliminary design of these modules, is described in Packet Radio Temporary Note 143, "Specification of Basic PRN Station Modules," which is completely written and in the

final stages of revision before release. We have begun coding of the "connection process" described therein which accepts packets from applications processes destined for the PRN and vice versa. It also handles retransmission and acknowledgements.

G. Cross-network debugger

Minor improvements to X-NET, our cross-network debugger, have been made as their importance was recognized and as time permitted. These include: halfword typeout mode; control-E to abort awaiting network replies, and control-Q to query their status; buffering of type-ahead while awaiting network replies; four times as many breakpoints; "end debug" clears process status and breakpoints, if any; and an improved "asynchronous reply" printout.

Time permitting, we wish to modify the cross-network debugger process in the PDP-11 so it allocates buffers from the ELF system's pool of buffers. This will be consistent with the buffer strategy of other ELF modules.

To ease the difficulty of debugging routines on a stand-alone PDP-11 (i.e., not running ELF or some other PDP-11 Operating System), a Cross-Net Debugger Server was written. This Server runs in Virtual mode with user programs running in User mode with memory management on. Besides protecting the Server from mis-behaving user routines, the Server is thus able to trap illegal memory references, illegal instructions, inadvertent modification of trap vectors, etc. This enables the programmer to obtain feedback on types of errors

that are very difficult to track down on a stand-alone PDP-11.

In conjunction with the above, a few trap instructions have been provided by the Cross-Net Debugger Server to give user programs full but controlled access to all device registers, trap vectors, and interrupt handling capabilities. Thus device drivers operating at high priority levels can be debugged.

This work has vastly eased the difficulty of debugging systems to be run on a stand-alone PDP-11.

H. Simple PDP-11 gateway

To test our gateway concepts, two steps have been taken. First, a simple gateway has been coded for the PDP-11. Second, two TCPs (transmission control programs) have been generated to run on PDP-10s. Each TCP simulates a network, and the PDP-11 will function as a gateway between them. We are now ready to debug the gateway code and confirm that the design is functional.

BOLT BERANEK AND NEWMAN INC

CONSULTING • DEVELOPMENT • RESEARCH

BBN Report No. 3093 June 1975

COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

Part II. Speech Compression

Quarterly Progress Report No. 2

1 March 1975 to 31 May 1975

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 2935. Contract No. MDA903-75-C-0180.

TABLE OF CONTENTS

I. INTRODUCTION

In the speech compression project, we have made progress on a number of issues in the last quarter. We considered transmission and implementation aspects of linear predictor gain, and made certain specific recommendations [1].

Next, by taking advantage of the differences in the spectral sensitivity of the various log area ratios, we developed an improved technique which employs unequal step sizes for quantizing these parameters.

By making several relatively simple modifications to our variable frame rate transmission scheme, we achieved average bit rates for continuous speech transmission as low as 1000 bits per second (bps). While the intelligibility of the transmitted speech in this case is still good, the speech quality has deteriorated somewhat from the 1500 bps system that we demonstrated earlier [2].

As another approach to speech data compression, we have used the well-known differential pulse code modulation (DPCM) method for removing the redundancy from the linear predictive (LP) system transmission parameters. The resulting compression system transmits speech at 2000 bps with speech quality essentially indistinguishable from that of the corresponding system with unquantized transmission parameters.

With the ultimate goal of developing extremely low bit-rate vocoder systems (about 500 bps), we briefly worked on an LP speech compression system which transmits only the first three formants.

We found that the particular formant-coded LP system we used does not have some of the problems encountered by the conventional formant vocoder.

Finally, for our SPS41-PDP11 signal processing system, the A/D and D/A hardware items were delivered, installed and checked out. Diagnostic routines were also developed for these components.

II. GAIN ISSUES

During the past quarter, we investigated three issues involving linear predictor gain parameter. The first issue is the choice of the gain parameter for transmission. The second issue considers the problems associated with implementing the speech signal energy as a multiplier at the output of the synthesizer filter instead of the more commonly used method of applying it at the filter input. The third issue is the treatment of cases for which speech signal energy has values less than 1 (or negative when considered in decibels). As we have discussed the first issue in detail in ARPA Network Speech Compression (NSC) Note No. 56 [1], we present below only a summary of the results.

As gain parameter, one can transmit either the energy of the speech signal, R_o, or the energy of the prediction error, E_p. These two quantities are of course related to each other:

$$E_p = R_o V_p , \qquad (1)$$

where V_p denotes the normalized error of the linear predictor. It can be shown that E_p has a smaller dynamic range and hence leads to a smaller quantization error than R_o. However, when transmitting E_p, a problem arises from the fact that the normalized error of the quantized predictor is different from the unquantized case. This causes error in the energy of the synthesized speech even when E_p is not quantized before transmission. This of course is not the case if we transmit R_o. Another consideration in deciding which

transmission parameter to use for gain is the type of synthesizer implementation. Regular filter realization (direct form or ladder structure) and normalized filter realization [3] are the two types used by the NSC group. The gain of the regular filter is equal to the square root of E_p, while the gain of the normalized filter is equal to the square root of R_o. Thus, for example, if the receiver employes the normalized filter, it is better to transmit R_o since transmitting E_p in this case requires computing the normalized error of the synthesizer filter and dividing with it the received E_p to obtain the normalized filter gain. Avoiding these extra operations may be desirable particularly for real-time implementation.

We have conducted a statistical error analysis using both R_o and E_p for transmission. Our findings indicated that, in general, it is better to use R_o for transmission than to use E_p. Such a choice is more strongly recommended when using the normalized filter. The results of this study also suggested a third alternative which is to transmit the product of R_o and the normalized error of the quantized predictor. This alternative seems attractive for the case when the regular filter realization is used [1].

The use of the normalized filter is recommended for implementation on the SPS-41 for many reasons, such as better round off noise and scaling properties, the availability of sine and cosine tables in the SPS-41, etc. Placing the gain multiplier, which is the square root of the speech signal energy as mentioned above, at the output of the normalized filter rather than at its

input serves to alleviate dynamic range problems. However, care has to be exercised in implementing the speech signal energy at the output of either the normalized filter or the regular filter. The difficulty, implied in the above statement, arises from the nonzero initial conditions of the filter. Whenever there is a relatively large change in speech signal energy from one frame to the next, say, of the order of 10 dB, then the synthesized speech is found to have signal amplitudes quite different from those of the original input speech. For example, in an unvoiced-voiced transition, the first voiced frame in the synthesized speech has relatively large signal amplitudes compared to the original speech. We have shown both experimentally and mathematically that this problem is due to the nonzero initial conditions of the filter. When listening to speech synthesized with speech signal energy implemented at the output of the synthesizer filter, we perceived these distortions in signal amplitudes as annoying "knock sounds". A solution to the problem, which we have found to be satisfactory, is to zero the initial conditions whenever the absolute frame-to-frame energy change exceeds a given threshold (about 12 dB). With this method, the distortions in signal amplitudes which caused the perception of "knock sounds" were eliminated. This was demonstrated at the recent NSC meeting (June 4-5, 1975).

In logarithmically quantizing speech signal energy we use a range of 0 to 45 dB. Any signal energy less than 0 dB is quantized as 0 dB. From synthesis experiments we found that this strategy of raising the energy from a negative dB value to 0 dB produced relatively large perceivable noise during stop sounds, pauses and

silences. This led us to quantize energy values less than or equal to 0 dB as a given negative dB. We found through listening tests that when we used a large negative dB value, the beginnings of certain speech sounds (e.g., [h], [n], [d]) were somewhat cut off. By experimentation, we found a value of -3 or -4 dB to be satisfactory.

In general, it is best to perform logarithmic encoding by table look-up. However, for encoding gain we have suggested a simpler method. The method is identical to encoding of numbers in terms of an exponent and a mantissa. The details of this method are given in NSC Note No. 56 [1].

III. QUANTIZATION OF LOG AREA RATIOS USING UNEQUAL STEP SIZES

We have shown that uniform quantization of log area ratios (LARs) is optimal in the sense of a minimax criterion [2,4]. This result was obtained using a spectral sensitivity characteristic of the reflection coefficients averaged over a number of speech sounds and over different reflection coefficients. In the past we used the same step size for quantizing all the LARs. However, when we averaged the spectral sensitivity of each reflection coefficient separately over a number of speech sounds, we found that while the sensitivity curves of the different reflection coefficients have the same general U-shape, they are located at different sensitivity levels. By taking advantage of these differences in sensitivity levels of the reflection coefficients or equivalently LARs, we have developed an improved quantization scheme that uses unequal step sizes for the different LARs.

Figures 1 and 2 depict averaged spectral sensitivity curves of individual LARs for respectively voiced and unvoiced speech sounds. Each figure has 12 sensitivity curves corresponding to 12 LARs and also an average of all these 12 sensitivity curves. In order to derive the step sizes for quantizing the LARs, first we need to transform, for each LAR, its sensitivity curve to one number which we shall call its average sensitivity level. For the ith LAR g_i, it is reasonable to define its average sensitivity level S_i as

$$S_i = \sum_{k=1}^{I_i} \frac{\delta S}{\delta g_i}(g_i=G_{ik})\, P_{ik}\,, \qquad (2)$$

where the range of g_i is represented by L_i equally spaced points

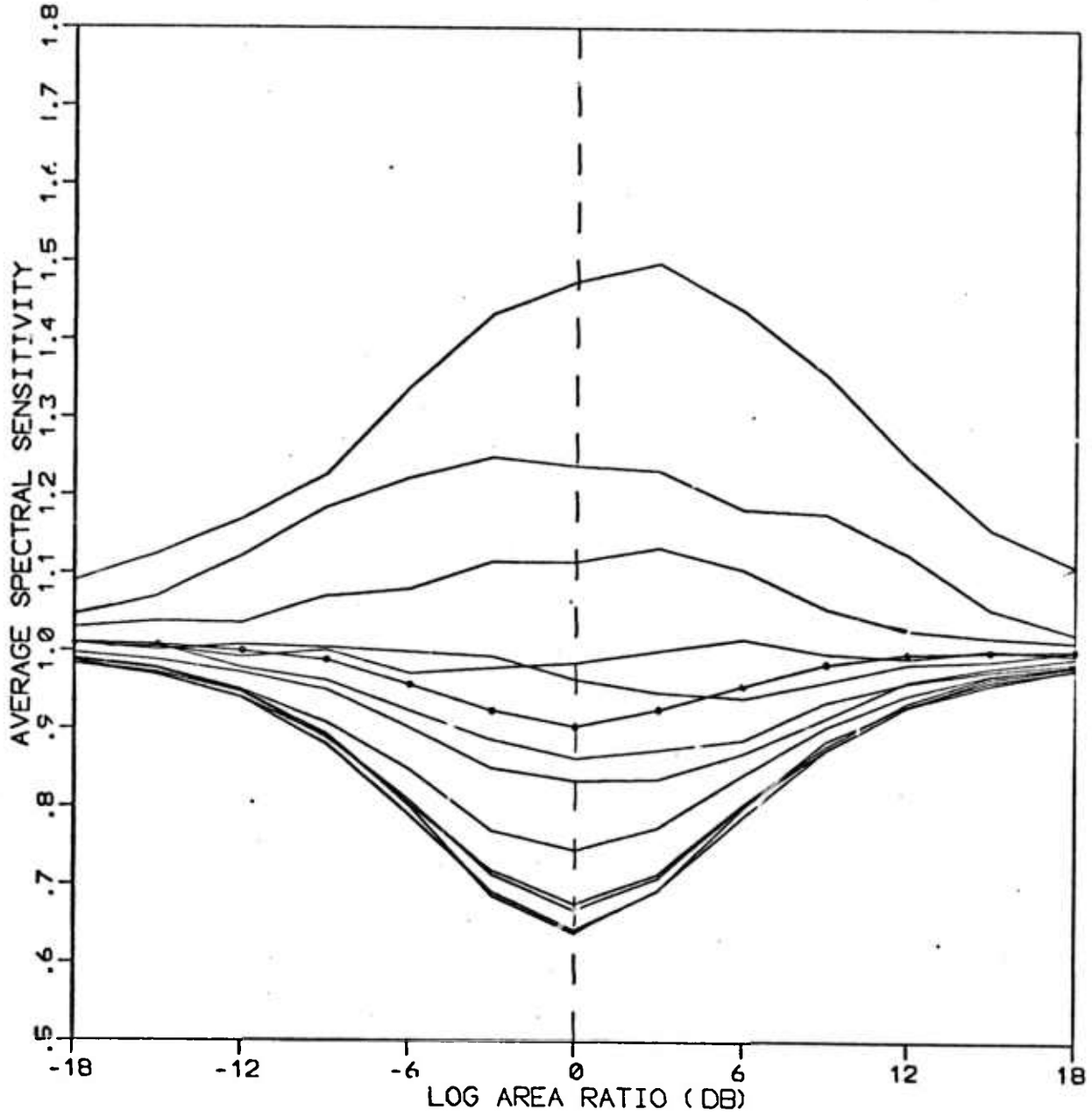

Figure 1. Spectral sensitivity curves for LARs of a 12th order linear predictor, averaged over voiced sounds only. The top curve corresponds to the first LAR; the bottom curve to the 12th LAR. Some sensitivity curves cross each other as shown. The average of the 12 sensitivity curves is drawn along circled points.

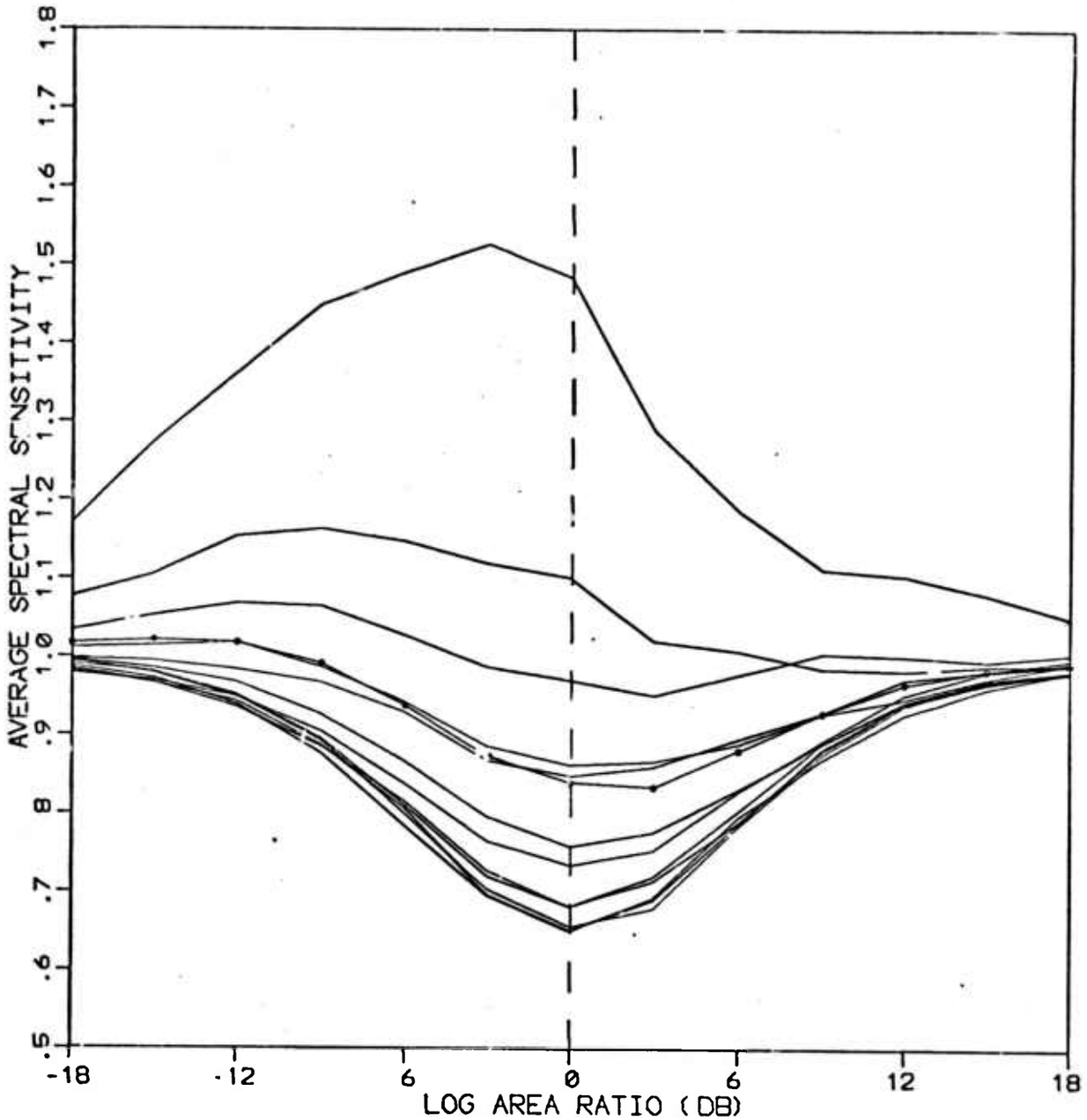

Figure 2. Spectral sensitivity curves for LARs of a 12th order linear predictor, averaged over unvoiced sounds only. The top curve corresponds to the first LAR; the bottom curve to the 12th LAR. Some sensitivity curves cross each other as shown. The average of the 12 sensitivity curves is drawn along circled points.

G_{ik}, $1 \le k \le L_i$; $\partial S/\partial g_i$ is the spectral sensitivity of g_i; P_{ik} is the probability of g_i taking the value G_{ik}. It is clear that S_i is approximately equal to the expected value of $\partial S/\partial g_i$ if L_i is sufficiently large. We used the data shown in Figs. 1 and 2 and the probability histograms of LARs that we prepared for Huffman coding work in computing the quantities S_i, $1 \le i \le p=11$, for both voiced and unvoiced cases.

Using the approach of optimal bit allocation strategy that we presented earlier [2,4], the number of quantization levels N_i and the step sizes ∂_i for the different LARs are computed as follows:

$$\delta_i = \frac{(g_i)_{max} - (g_i)_{min}}{N_i}, \quad 1 \le i \le p,$$

$$K_i = [(g_i)_{max} - (g_i)_{min}]\, S_i, \quad 1 \le i \le p, \qquad (3)$$

$$N_1 = K_1 \frac{2^M}{\prod_{i=1}^{p} K_i}^{\frac{1}{p}},$$

$$N_i = \frac{K_i}{K_1} N_1, \quad 2 \le i \le p,$$

where $(g_i)_{max}$ and $(g_i)_{min}$ are the upper and lower bounds on g_i, and M is the number of bits for quantizing all the p LARs. To compare unequal step size quantization with equal step size quantization, we have listed in Table 1 the numbers of quantization levels for these two methods with the same total number of bits and considering voiced and unvoiced cases separately. As expected, relative to the equal step size method, the unequal step size method places more emphasis on the first three LARs by alloting more levels to them. Synthesis experiments showed that use of the unequal step size quantization method produced better quality speech. The perceived quantization noise in the synthesized speech was reduced noticeably for the low bit rate system (1000 bps) discussed in the next section.

It should be noted that for real-time implementation, while the equal step size method requires only one coding table and one decoding table, the unequal step size method in general requires p coding tables and p decoding tables.

Table 1. Quantization Levels

Coeff. #	VOICED (43 BITS)		UNVOICED (41 BITS)	
	Equal Step (1dB)	Unequal Step	Equal Step (1dB)	Unequal Step
1	28	43	29	51
2	22	31	21	28
3	19	24	14	18
4	15	17	13	15
5	14	15	10	11
6	13	13	9	9
7	13	12	12	11
8	14	12	11	10
9	12	10	10	9
10	11	9	10	9
11	9	7	9	8

IV. 1000 BITS/SECOND SPEECH TRANSMISSION SYSTEM

Last year we demonstrated a good quality speech transmission system operating at average rates of 1500 bps for continuous speech (that is, no explicit detection of silences was made; silences were treated in the same way as any other speech sound) [2]. One of our goals in the last quarter was to modify this system so as to lower the average bit rate to 1000 bps, also for continuous speech, and to observe the type of speech quality produced by the resulting system.

Before we discuss the details of the 1000 bps system, we may recall some of the main features of the 1500 bps system. Analysis of the 10 kHz sampled speech was done every 10 msec using a variable number of poles (order of the linear predictor). The order was changed in accordance with the incoming speech signal characteristics, the maximum order used being equal to 11. Pitch and gain were transmitted at a fixed rate of 50 frames/sec, while a variable frame rate scheme was used to transmit the LARs only when speech characteristics changed sufficiently since the last transmission. The latter scheme employed a log likelihood ratio test with an error threshold of 1 dB for deciding when to transmit. This yielded an average rate of 37 frames/sec for the LARs. The interval between successive transmissions varied between 10 and 80 msec. Finally, Huffman coding was used to encode the quantized LARs and the frame-to-frame changes in the quantized values of pitch and gain. Other details of the 1500 bps system are presented in [2].

To achieve 1000 bps transmission, we made several modifications to the above system. First, pitch and gain were transmitted at a variable frame rate along with the LARs. This method will certainly cause problems if a transmission interval contains one or more transitions between voiced and unvoiced sounds. To overcome this problem, in addition to transmitting pitch and gain along with the LARs, we also transmitted two sets of pitch and gain values for every transition occurring within a transmission interval. For example, for a voiced-unvoiced transition, pitch and gain of the last voiced frame and gain of the first unvoiced frame would be transmitted. Such a variable frame rate transmission of pitch and gain was found to be quite satisfactory in terms of the resulting speech quality. This scheme by itself produced a saving of about 150 bps relative to the 1500 bps system.

Next, we experimented with the case where analysis was done only every 20 msec, so that transmission occurred at multiples of 20 msec. The maximum transmission interval was still kept as 80 msec. Clearly, for a given average bit rate, the case of 20 msec analysis will permit a lower threshold to be used in the log likelihood ratio test (of the variable frame rate scheme) than the case of 10 msec analysis. However, a number of synthesis experiments indicated that the 10 msec analysis method actually produced a slightly better overall speech quality, improvements mainly occurring at instances of rapid speech transitions.

A number of changes were also made to the variable frame rate scheme so as to lower the average frame rate of data transmission but still maintain the quality of the transmitted speech at a reasonable level. Clearly, the frame rate can be lowered to any value simply by increasing the threshold in the log likelihood ratio test by a sufficiently large amount. However, the quality of the synthesized speech in this case would be rather poor. One simple change that did not cause any perceivable distortion was to use a slightly higher threshold (about 15-20% higher) for unvoiced sounds than for voiced sounds. When a transmission interval contained a transition between voiced and unvoiced sounds, the lower threshold was always employed to encourage a transmission. Lowering the average bit rate to 1000 bps required an increase in the thresholds to such large values that a number of problems occurred in the synthesized speech. For example, in a transition between a vowel and a nasal, only the data for the vowel was transmitted. This produced lack of nasalization that could be clearly perceived. We therefore improved the variable frame rate scheme as follows. A double threshold strategy was used for both voiced and unvoiced sounds. When the log likelihood ratio error measure between a current data frame and the previously transmitted data frame did not exceed the first threshold (smaller one), the current data frame was not transmitted. If it exceeded the first threshold, it was then compared against a second threshold (larger one). If the second threshold was also exceeded, the data frame immediately preceding the current frame (but excluding the last transmitted frame) was transmitted; in all other cases, the current data frame was

transmitted. By carefully selecting the values of the two thresholds, the problems referred to above were overcome considerably. With 10 msec analysis, these thresholds were 4.25 and 6 dB for voiced sounds, and 5 and 6.75 dB for unvoiced sounds. This choice resulted in an average data rate of 22 frames/sec.

We used about the same number of bits per data frame as in the 1500 bps system. However, we employed the unequal quantization method discussed in Section III. Another added feature was the use of the optimal interpolation scheme [5]. The improvement in speech quality due to this scheme was well worth the slight increase of about 50 bps in the bit rate.

We demonstrated the 1000 bps system described above at the June NSC meeting. Informal listening tests showed that while the intelligibility of the speech transmitted at 1000 bps was still good, the speech quality, however, deteriorated somewhat relative to the 1500 bps system. In a comparative experiment using the triplets of the 1000 bps speech, the 1500 bps speech and the unprocessed original speech, the biggest drop in quality occurred, as would be expected, in going from the unprocessed speech to the 1500 bps speech.

V. CODING OF TRANSMISSION PARAMETERS USING DIFFERENTIAL PULSE CODE MODULATION

DPCM is a well-known method for quantizing signals which exhibit high correlation between successive samples. This method has been widely used for coding speech signals [6,7]. Following a recent work [8], we have used the DPCM method for coding the LARs, pitch and gain. Each of these transmission parameters is considered as a discrete-time signal with time instants given by the frame number. DPCM is applied to each of these signals independently of others. For p=12, then, 14 DPCM coders and decoders may be required. However, hardware implementation of these is relatively easy and inexpensive [7].

A basic DPCM coder-decoder system is depicted in Fig. 3. At any time instant i, a prediction f_i is made of the input signal x_i. The difference x_i-f_i, or the prediction error, is quantized as y_i which is then encoded and transmitted. At the receiver, the decoded sample is y_i' is added to its predicted value f_i' to generate $\hat{x}_i'$ as an approximation to the input sample x_i. The predictors at the transmitter and the receiver are the same. The feedback structure around the quantizer at the transmitter ensures that quantization errors do not accumulate in forming $\hat{x}_i$ or $\hat{x}_i'$. For the system shown in Fig. 3, the overall system error $\hat{x}_i'-x_i$ is, in the absence of channel errors, equal to the corresponding quantization error z_i-y_i.

In our investigation, the predictor was taken as a unit delay, i.e., $f_i=x_{i-1}$. After reference [9] we assumed a gamma distribution for the difference signal z_i. This reference has also tabulated the normalized step sizes for optimum nonuniform quantization of the

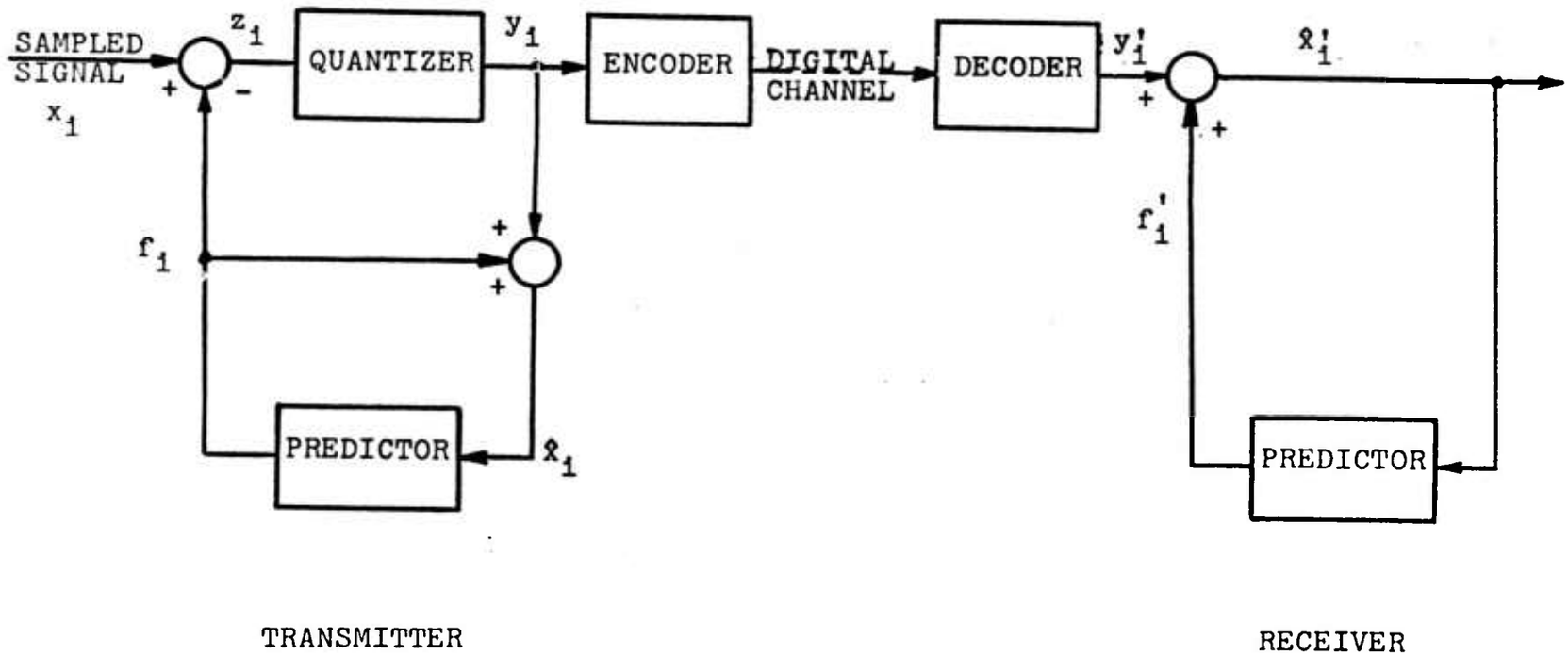

Figure 3 Basic DPCM method.

difference signal (optimum in the sense of minimum mean square quantization error). We ran DPCM experiments on a number of unpreemphasized speech utterances analyzed every 20 msec, and computed the standard deviation of the difference signal for each of the 14 transmission parameters. Actual quantization step sizes were obtained by multiplying the normalized step sizes with the corresponding standard deviation.

We applied the DPCM method for coding the 14 transmission parameters (p=12 LARs, pitch and gain) extracted at a fixed rate of 50 frames/sec from 10 kHz sampled and unpreemphasized speech. One bit was transmitted with every frame to code its voiced/unvoiced status. For initializing the DPCM process, the transmission parameters were quantized in the usual manner (linear quantization of LARs, logarithmic quantization of pitch and gain) and transmitted once at the very beginning of speech input and then once at every transition between voiced and unvoiced frames. Clearly, for unvoiced frames, there is no need to quantize pitch using DPCM.

In the first set of experiments, we employed the following bit allocation for quantizing the various difference signals: Pitch, 4 bits; gain, 3 bits; LARs, 4, 4, 4, 4, 3, 3, 2, 2, 2, 1, 1, 1 bits. This led to an average bit rate of about 2000 bps. Informal listening tests were conducted using loudspeakers as well as earphones to compare the 2000 bps synthesized speech with the speech synthesized using the same 20 msec (50 frames/sec) but <u>unquantized</u> (i.e., represented to 36-bit accuracy) parameter data. For this test, we employed the six utterances used in the speech quality

evaluation project. It was found that the 2000 bps system produced speech quality essentially indistinguishable from that of the unquantized case. This result by itself is rather important, and was demonstrated at the June NSC meeting. However, speech obtained from the 20 msec unquantized system is not high quality speech. In fact, our 1500 bps variable frame rate system was found to produce speech at a slightly better overall quality level than the above unquantized system.

In another set of experiments, we used a fewer number of bits/frame for DPCM quantization: Pitch, 3 bits; gain, 2 bits; LARs, 3, 3, 2 bits and all the rest 1 bit each. This bit allocation was found to produce bit rates of about 1150 bps. The resulting speech quality, however, was inferior to that of our 1000 bps system described in Section IV.

In the course of these experiments, we observed an important consideration for the DPCM system. For the results reported above, we averaged standard deviation data over a number of speech utterances and used the averages for DPCM quantization. The 1150 bps DPCM system produced considerably better quality speech when we used standard deviation data averaged only over the particular utterance (typically 2.5 seconds long) being synthesized. Of course, such an approach is an off-line method in the sense that one needs to compute first the standard deviation data for the whole utterance before starting the DPCM quantization process. The point of mentioning this observation is that it naturally suggests the use of a suitable adaptive DPCM method [7].

In conclusion, the DPCM method offers another approach to achieve low bit rate speech transmission. The variation of the actual bit-rates about the average value was found to be relatively small, suggesting its usefulness in applications where a fixed-rate transmission is required. Another advantage, mentioned above, is that hardware implementation of DPCM coder and decoder is relatively simple and inexpensive.

VI. LINEAR PREDICTIVE FORMANT VOCODER

It has been known for some time that formant vocoders enable speech transmission at very low bit rates (about 500 bps). One requires of these systems an acceptable level of speech intelligibility but not necessarily retention of naturalness or speaker characteristics. Such low-bit rate systems are of interest in some applications. Speech transmission through an underwater channel is a good example.

In the last quarter we conducted a preliminary experiment simulating a formant vocoder within our LP system format. Formants were generated from LP analysis data. The formant synthesizer was implemented not using resonators as in conventional formant vocoders, but employing the canonical or direct form realization of the LP all-pole filter. The predictor coefficients of the all-pole filter were computed from the received formant data. It is this difference in synthesizer implementation which has enabled our LP formant vocoder to overcome some of the problems encountered by its predecessors. The LP formant vocoder can accommodate variable number of formants in adjacent frames without causing any undesirable transients. Incorrect identification of formants, which in practice can occasionally happen due to imperfect formant tracking, produces less degradation in the quality of synthesized speech for the LP formant vocoder than for its conventional counterparts. A third advantage stems from the result we reported last year that the parameters of the LP synthesizer filter can be updated time-synchronously without introducing any transients. It

is well-known that such transients occur if one updates the parameters of the resonators time-synchronously.

In the preliminary experiment, we employed the formant data already computed in our Speech Understanding Project. There, a 14-pole LP analysis was done every 10 msec on speech sampled at 10 kHz and preemphasized using a 50 Hz first-order filter. The formant tracker used in that project then extracted, every 10 msec, up to a maximum of 3 formants in the frequency range 0-3100 Hz. For unvoiced sounds, often only two formants were determined. Gain and pitch were also computed every 10 msec. For the purposes of the preliminary experiment, we did not quantize any of these analysis parameters. The receiver thus had a variable order LP synthesizer. The synthesized speech was found to be quite intelligible except for the following type of problem: [s] was often perceived as [sh]. The reason for this problem is that [s] has significant energy concentration above 3.1 kHz unlike [sh] and that we essentially low-pass filtered speech at 3.1 kHz by considering only those formants below this frequency.

VII. REAL TIME SYSTEM

During the past quarter we have continued to check out our system as hardware modifications have been made to it by SPS. A major modification involved the installation of new cards to repair design errors in the dual-port memory interface. A second major modification was the installation of our dual channel analog to digital and digital to analog converter system. We developed diagnostics for this system and monitored its installation. Another addition to our SPS41-PDP11 system is our IMP11A network interface, which has arrived and has been installed and tested.

VIII. PUBLICATIONS AND PRESENTATIONS

During the last quarter, NSC Note No. 56 was written on some issues involving linear predictor gain [1]. A paper entitled "Optimal Linear Interpolation in Linear Predictive Vocoders" was presented at the 89th meeting of the Acoustical Society of America, and has been published as BBN Technical Report No. 3065 and NSC Note No. 59. A survey paper on linear prediction was published in the Proceedings of IEEE [10]. Two other papers, one on spectral linear prediction [11] and the other on quantization of linear predictor transmission parameters [4] were published in the Transactions of IEEE on Acoustics, Speech and Signal Processing.

REFERENCES

1. J. Makhoul and L. Cosell, "Nothing to Lose, But Lots to Gain," NSC Note No. 56, March 1975.

2. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Final Report, Vol. II, Speech Compression Research at BBN, Report No. 2976, Bolt Beranek and Newman Inc. Cambridge, Mass., Dec. 1974.

3. A.H. Gray and J.D. Markel, "A Normalized Digital Filter Structure," NSC Note No. 19, Feb. 1974.

4. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, pp. 309-321, June 1975.

5. R. Viswanathan, J. Makhoul and W. Russell, "Optimal Interpolation in Linear Predictive Vocoders," Report No. 3065, Bolt Beranek and Newman Inc., Cambridge, Mass., April 1975 (also NSC Note No. 59).

6. R.A. McDonald, "Signal-to-Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems - Particular Applications to Voice Signals," Bell Sys. Tech. Journal, pp. 1123-1151, Sept. 1966.

7. N.S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers," Proc. IEEE, Vol. 62, pp. 611-632, May 1974.

8. M.R. Sambur, "Efficient LPC Vocoder," presented at the 89th Meeting of the Acoustical Society of America, Austin, Texas, April 7-11, 1975.

9. M.D. Paez and T.H.Glisson, "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems," IEEE Trans. Communications, pp. 225-230, April 1972.

10. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, Vol. 63, pp. 561-580, April 1975.

11. J. Makhoul, "Spectral Linear Prediction: Properties and Applications," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, pp. 283-296, June 1975.

BOLT BERANEK AND NEWMAN INC

CONSULTING · DEVELOPMENT · RESEARCH

BBN Report No. 3093

June 1975

COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

Part III. Vocoded-Speech Evaluation

Quarterly Progress Report No. 2

1 March 1975 to 31 May 1975

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 2935, Contract No. MDA903-75-C-0180.

III. VOCODED-SPEECH QUALITY EVALUATION

Table of Contents

III. VOCODED-SPEECH QUALITY EVALUATION

1. INTRODUCTION

The evaluation procedures that were reported in the preceding Quarterly Progress Report required listeners to rank-order a set of LPC vocoding systems according to the quality of speech they produced. Averaging such rank orders across listeners, and across different types of speech material, spoken by different speakers, can produce data that are clearly helpful for selecting the best of a set of candidate systems. Inspection of the data showed, however, that the rank orders were considerably affected by both the speech materials and the speaker, as we had anticipated. This suggests that the vocoding systems differed from each other in more than one respect and that finer analysis might lead to an understanding of which attributes of a sample of speech determine its position in the rank orderings. This understanding might well have diagnostic utility, by highlighting those attributes of a particular system that are most responsible for degrading the quality of speech processed through it. In other words, although it is of interest to know that the speech produced by System A is in a general sense preferred to that produced by System B, it would be of even greater interest, at least to developers of speech-processing procedures, to know what it is about the speech produced by a given system that makes it better or worse than that produced by another.

There are two ways to attack this problem. One is to assume that the inconsistencies in the rank-order data are not due only to variability or error on the part of the listeners, as the averaging of rank orders would imply, but, rather, that they reflect structure in the data that is discarded by averaging, but which could be

recovered by more sophisticated analysis. This approach, which most commonly uses multidimensional scaling (MDS) procedures, attempts to infer from the rank orders the basis on which those rank orders were produced.

The second possible approach is to assume from the outset that the speech from different vocoders may vary along several perceptually independent dimensions, and attempt to determine the dimensions directly, and then measure speech quality explicitly on each dimension. These data too can be analyzed by multidimensional scaling procedures, to extract the optimal dimensions, which are usually not quite identical to those found empirically.

We have performed some exploratory experiments with both of the foregoing approaches. We will discuss the MDS analysis of the rank-order data first, since it involves only the data obtained with the listening procedure already described in the preceding QPR (Section 6.2).

2. MULTIDIMENSIONAL SCALING

The aim of all psychological scaling procedures is to assign numbers to perceived properties of stimuli in such a way that the psychological magnitude of a stimulus can be predicted from some other, more easily measured, property, such as some physical property. Thus in scaling the loudness of sounds, one attempts to "place" the various sounds on a line in such a way that when one sound is twice as loud as another, it appears at twice the scale value on the line. For simple attributes such as loudness or brightness, this procedure produces very orderly data. The distances between the stimuli on the scale represent the psychological distances between the stimuli as perceived, or their similarity. Some sorts of data, however, cannot be fitted by such

a model. For example, if the psychological distances between stimuli A and B, B and C, and A and C are all equal and non-zero, the three stimuli cannot be placed on a one-dimensional scale without doing violence to the measured distances. But they can easily be accommodated in a two-dimensional space. Multidimensional scaling is a generalization to n-dimensional space of the procedures already developed for one-dimensional space, that is, position on a line. Two additional problems are introduced by the generalization: how should distance be measured within the space, and how many dimensions are required to account adequately for a given set of data? The space is usually assumed to be Euclidean, so that the distance between two points is the square root of the sum of the squares of the differences on each separate dimension. The assumption is made partly for computational convenience, and partly because it seems to be empirically justified, since it has led to several intuitively reasonable solutions to scaling problems (see Carroll & Wish, 1974, for a detailed discussion). The decision on how many dimensions are needed for the model to capture adequately a given set of data is usually made by performing analyses in several dimensionalities, and using graphs of goodness-of-fit against dimensionality to decide how many dimensions are appropriate. With speech-evaluation data, precision typically approaches an asymptotic level fairly quickly, and the gains diminish greatly as n is increased beyond 2 or 3 (McDermott, 1969; McGee, 1964, 1965). It is also the case, as McDermott (1969) points out, that reliability tends to decrease as dimensionality is increased.

Several points are worth emphasizing with respect to the solution space generated by an MDS analysis. First, it is a perceptual, or subjective, space; the axes correspond to subjective factors. Second, the analysis itself does not identify what those factors are; it only indicates how well n of them can account for

the data. One can sometimes make a reasonable guess concerning what one or more of the axes represent by simply noting the way the stimuli are distributed throughout the space, but this is not always possible. Third, the subjective factors represented by the axes may or may not have physical analogs; that is to say, it may or may not be possible to associate the axes of the subjective space with objective properties of the stimuli. Fourth, since distance in a Euclidean space is invariant under rotation of the axes, the axes of the solution may be rotated if this would lead to a better interpretation of the space. (This ceases to be true in some models, such as INDSCAL, which weight each of the axes idiosyncratically for each subject.)

There are several ways in which rating of quality, or preferences, can be represented within the space containing the stimuli. One is to represent each subject by a vector, such that his preference score for each stimulus corresponds to the projection of that stimuli onto the vector, The vector model is a special case of a more general method for representing preferences. In the more general method, each subject, as well as each stimulus, is represented in the space by a point, and a subject's preferences are related to the distances from "his" point to each of the stimuli, with the nearest being the most preferred. In some ways, this more general model is preferable to the vector model, since it does not entail the sometimes counter-intuitive assumption that if so much of a particular attribute is good, then more of it must be better. For example (Carroll, 1972) although a cup of tea with one spoon of sugar may be preferred to a cup without sugar, it does not follow that a cup with ten spoons of sugar will be preferred to one with eight! It is not clear whether such inversions occur in perceptual dimensions relevant to speech quality, but they may. For example, natural voices show some jitter in the period of the fundamental frequency (Lieberman, 1963). This jitter

is absent in most synthetic speech, which may in part account for its machine-like quality. On the other hand, abnormally large jitter is indicative of pathology of the larynx. Thus, small amounts of jitter may sound more natural than either smaller or larger amounts. Despite these difficulties, the vector model has the advantage that it is easy to understand and interpret. In particular, analyses can be performed which represent as vectors either the individual listeners (subjects), or the different speakers, or the different sentence materials, or any desired combinations of these. Projections of the stimulus points on a vector provide an approximation (the best possible, given the constraints of the analysis) of the performance of the associated subject, or under the associated condition. The position of a vector vis-a-vis the coordinates of the space is an indication of the relative importance (weighting) of each of the attributes in determining the preferences of that subject, or under that condition. If, for example, a vector were coincident with one of the axes of the space, the implication would be that preferences were determined solely by the attribute represented by that axis. If the angular distance between a vector and two axes were equal, the implication would be that two attributes were equally weighted.

A critical assumption underlying the MDS approach is that the same subjective stimulus attributes form the basis for the judgments for all subjects. The use of vectors to represent individual subjects allows for differential weighting of those attributes, but the attributes themselves are assumed to be invariant.

2.1 Application of MDS to Rank-Order Data

It was noted in the preceding QPR that two MDS computer programs, MDPREF and INDSCAL, had been acquired from Bell Laboratories

and were being modified to run on Tenex. The modifications of both programs have been completed, and we have analyzed our data from the rank-ordering experiment (see preceding QPR, Section 6.2) with MDPREF. (The modification of INDSCAL has just been completed, but the data have not yet been alalyzed with that program.) The results demonstrate quite clearly that the ordering of systems on a preference scale may depend strongly on both the characteristics of the speaker's voice and the nature of the sentence material used in testing.

The MDPREF analysis was performed in 1, 2, and 3 dimensions. We will concentrate in this report on the results of the two-dimensional analysis, inasmuch as two dimensions were sufficient to account for almost 98% of the variance. We have already noted that other attempts to apply multidimensional scaling methods to the analysis of speech-quality judgments have typically found no more than two or three perceptual dimensions underlying quality (McDermott, 1972; McGee, 1964, 1965).

Figure 1 shows the results of an analysis that pooled data across listeners and across speakers, but preserved the variability introduced by the sentence materials. A total of 144 judgments (4 listeners, 6 talkers, 6 sentences) influence the placing of each vocoder system, which is represented by a point. The points and the vectors are so placed as to optimize the agreement between the observed preferences, and the preferences predicted by the model. That is, the model achieves a best fit simultaneously of the fourteen systems, assessed by all six sentences. It can be seen that the ranks of some systems are strongly affected by the choice of sentence. Figure 2 illustrates this more clearly for sentence 1 (Why were you away a year, Roy?) and sentence 4 (Which tea party did Baker go to?). Figure 2 shows only the projections

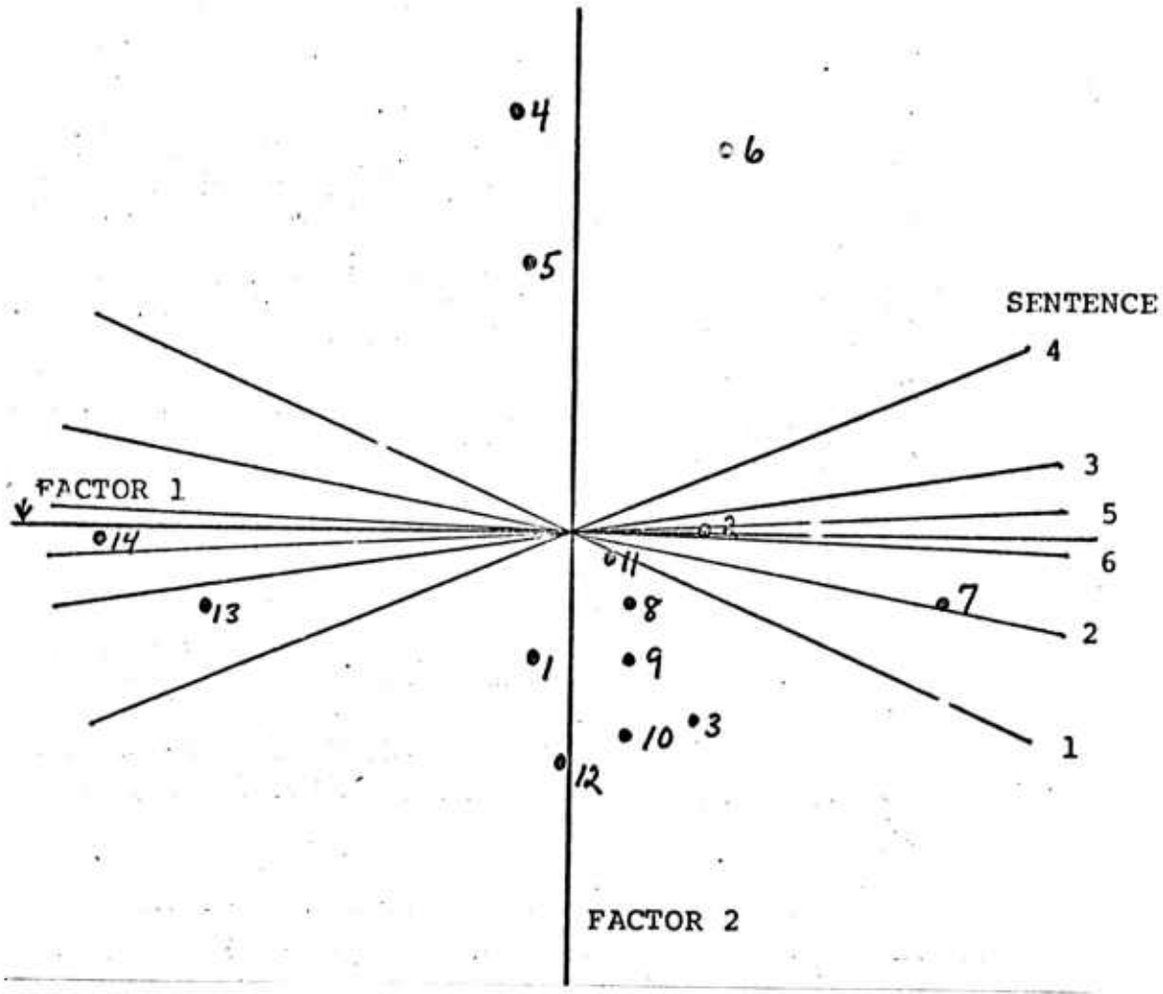

Fig. 1. Results of MDPREF 2-dimensional analysis for effects of sentence materials.

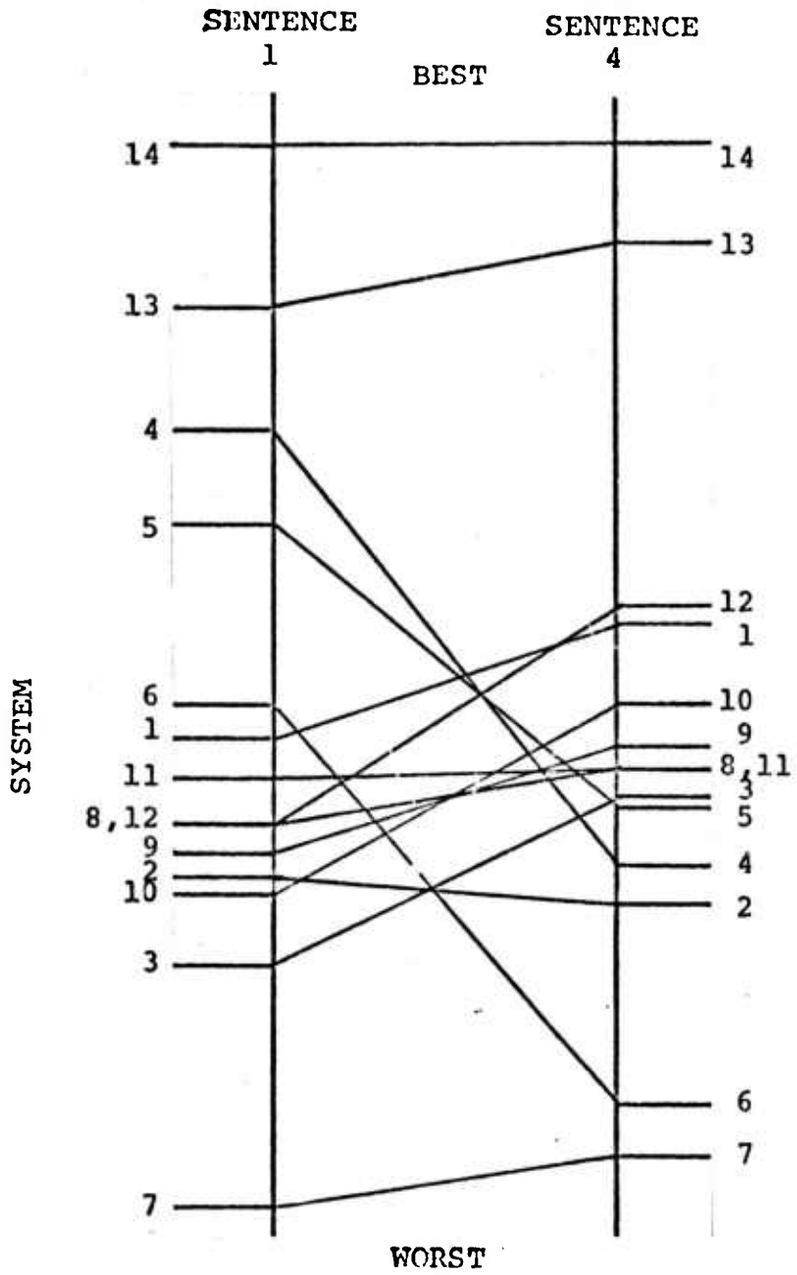

Fig. 2. Projections of the points representing systems in Fig. 1 onto the vectors representing sentences 1 (Why were you away a year, Roy?) and 4 (Which tea party did Baker go to?)

of the points (systems) on the two vectors, with lines joining corresponding systems. Some systems (e.g., 2, 7, 11, 13, 14) preserve their positions relatively independently of the sentence material, whereas others (e.g., 3, 4, 5, 6, 10, 12) shift considerably.

Figure 3 shows the results of a similar two-dimensional analysis that preserved only the variability introduced by the talkers. Figure 4 shows the same analysis, except in this case each talker-sentence combination was treated separately. Only the ends of the vectors are drawn, so as not to clutter the figure.

To interpret these results it is necessary to consider the properties of the sentence materials, the voice characteristics of the talkers, and the parameters of the systems used in the test. These data are shown in Tables 1 through 4. Perhaps the importance of talker and sentence effects is best demonstrated by focusing on the extreme vectors of Fig. 4. The vector that runs top-left to bottom-right represents the combination of talker RS and sentence 1. The one running from bottom-left to top-right represents talker AR and sentence 4. The projections of the points onto these two vectors are shown in Fig. 5. Several systems proved to be quite susceptible to combined talker-sentence effects. The most striking example is system 6, whose output was ranked nearly as high as the unquantized speech (13) with the talker-sentence combination RS-1, and the very worst with AR-4. The result is understandable after inspecting Tables 1 through 4. Sentence 1 is voiced throughout, and contains only vowels and /w, r, y/. These sounds are all characterized by slow rates of change of both spectrum and of envelope. It is the "smoothest" sentence, totally free of abrupt changes. Sentence 4, on the other hand, is the least smooth. All the consonants (except an

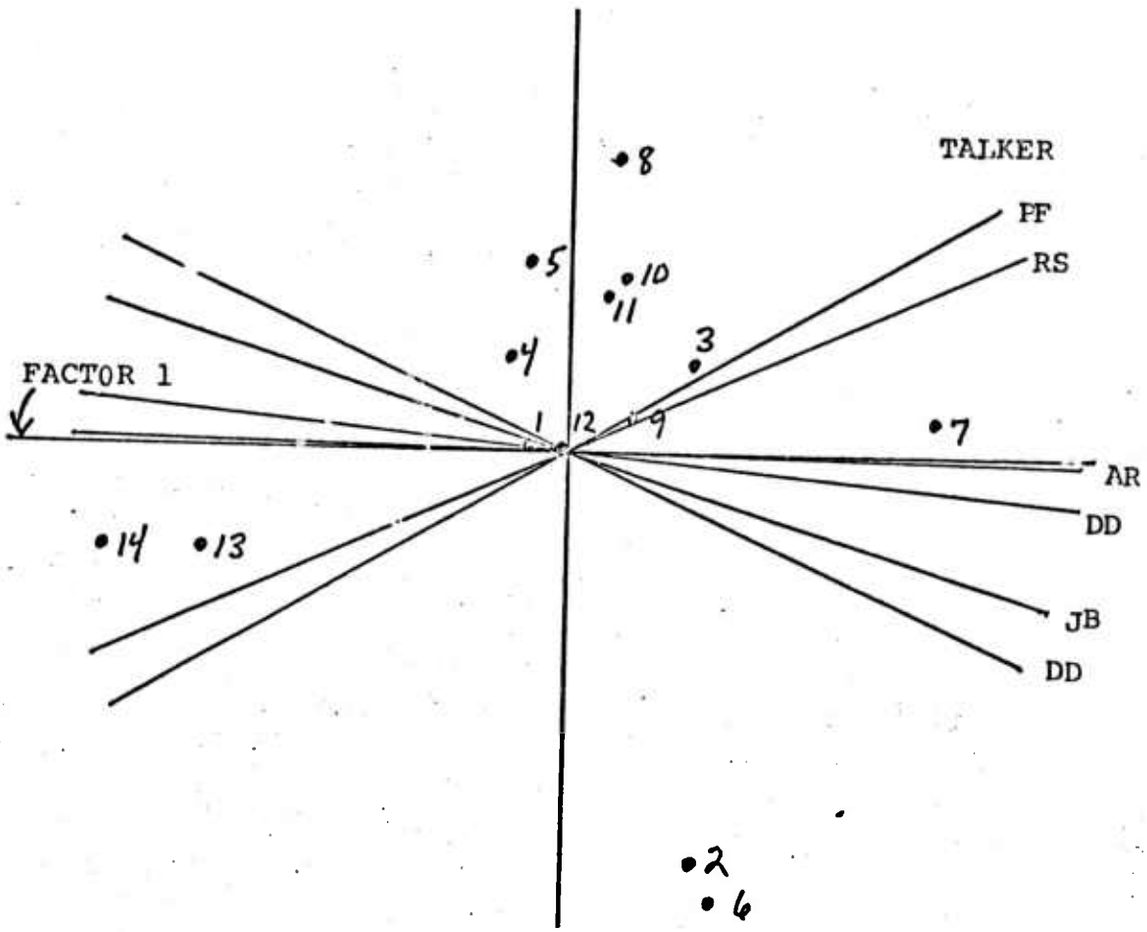

Fig. 3. Results of MDPREF 2-dimensional analysis for effects of talker.

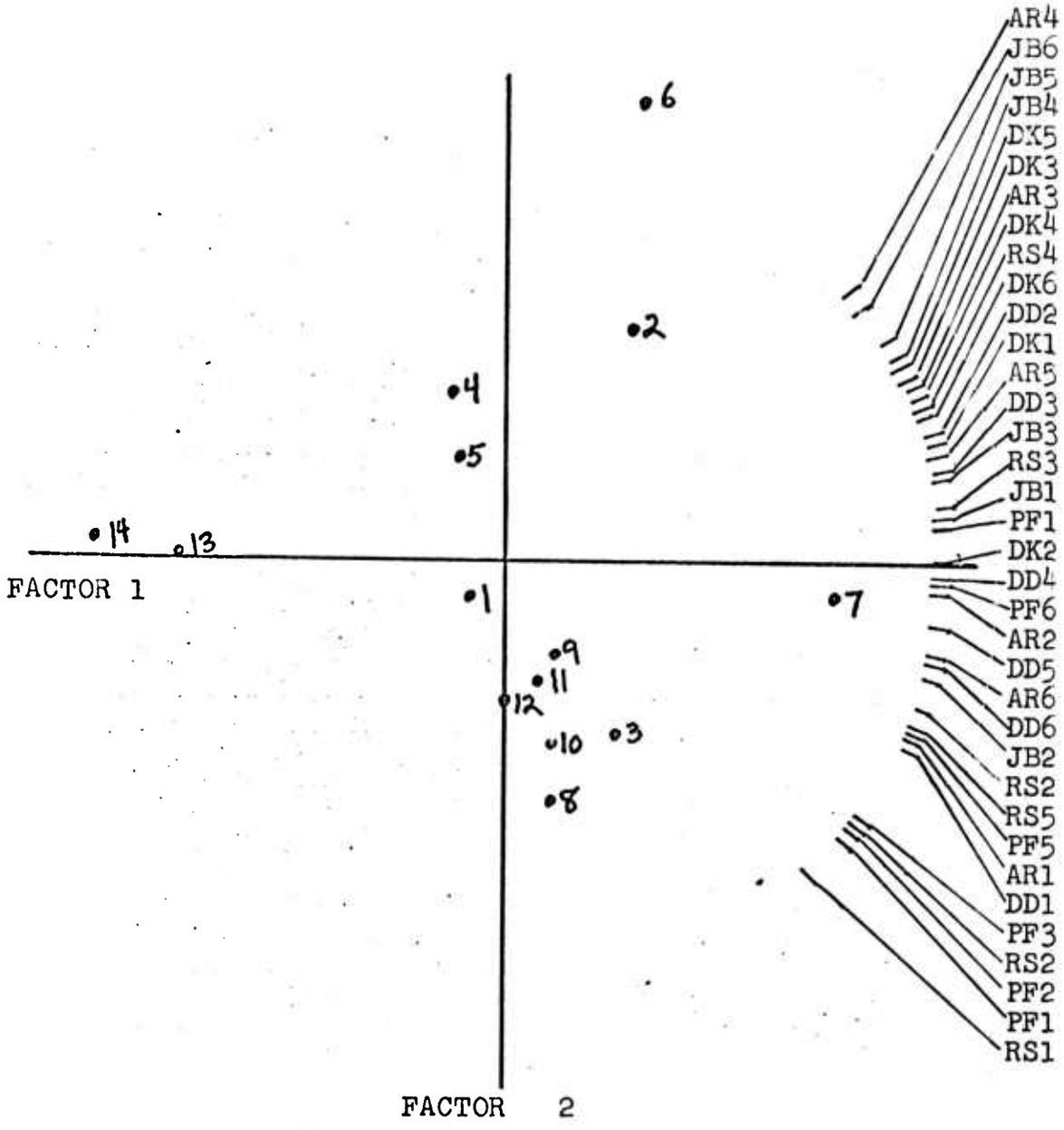

Fig. 4. Results of MDPREF 2-dimensionsl analysis for effects of talker-sentence combinations

Table 1. Sentences used in listening tests. For detailed phonetic breaddown, see Table 3.

1. Why were you away a year, Roy?
2. Nanny may know my meaning.
3. His vicious father has seizures.
4. Which tea party did Baker go to?
5. The little blankets lay around on the floor.
6. The trouble with swimming is that you can drown.

- -

Table 2. Speech characteristics of talkers used to record sentences. See QPR #1 for explanation of measurements.

SELECTED TALKERS

	Fo (Hz)	Rank (1-10)	Nasal (dB)	Rank (1-10)	Duration (Sec.)	Rank (1-10)
Males						
2	95	3	17.3	9	2.05	4.5
4	139	8	16.4	6	2.00	3
6*	118	6	14.3	1	2.20	6.5
Females						
3	167	2	15.2	2	2.00	1
18	232	9	17.5	6	2.70	9.5
19	209	6	17.0	4.5	2.40	4.5

*highly inflected

Table 3. Characteristics of sentences used for listening tests.

				sentences	1	2	3	4	5	6	Σ
Vowels	i	IY	beat		1	1	1	1			4
	I	IH	bit			2	2	3	2	4	13
	ɛ	EH	bet								0
	æ	AE	bat			1	1		1		3
	ɑ	AA	bob				1	1	1		3
	ɔ	AO	bought						1		1
	o	OW									0
	ʊ	UH	book								0
	u	UW	boot		1			1		1	3
	ʌ	AH	but							1	1
	ə	AX	(a)bout		2		1		2	3	8
	ɝ	ER	bird		2		2	2	2		8
Diphthongs	eɪ	EY	bait		1	1		1	1		4
	aɪ	AY	bite		1	1					2
	ɔɪ	OY	boy		1						1
	aʊ	AW	down						1	1	2
	oʊ	OW	boat			1		1			2
Semivowels	l	L							5	1	6
	r	R			1				1	2	4
	w	W			3			1		2	6
	y	Y			2					1	3
Unvoiced Stops	p	P						1			1
	t	T						3	1	1(1)	5(1)
	k	K						1	1	1	3
Voiced Stops	b	B						1	1	1	3
	d	D						2	1(1)	1	4(1)
	g	G						1			1
Nasals	m	M				3				1	4
	n	N				4			2	2	8
	ŋ	NX	sing			1			1	1	3
Unvoiced Fricatives	f	F					1		1		2
	θ	TH	(th)ink							1	1
	s	S					2		1	1	4
	ʃ	SH	(sh)ip				1				1
Voiced Fricatives	v	V					1				1
	ð	DT	(th)it				1		2	2	5
	z	Z					3			1	4
	ʒ	ZH	sei(z)ure				1				1
Affric + H	h	HH					2				2
	ʧ	CH	church					1		(1)	1(1)
	ʤ	JH	judge								0

Table 4. List of parameters of systems in listening tests.

System No	No of Poles	Frame Size (msec)	Var. Rate Threshold dB.	Quantization Step Size dB.	Bit - Rate Expected	Bit - Rate Obtained
1	12	20		1.0	2650	2630
2	10	20		0.6	2650	2633
3	14	20		1.4	2700	2681
4	12	25		0.45	2640	2610
5	14	25		0.7	2640	2612
6	10	25		0.2	2680	2652
7	10	15		1.75	2666	2618
8	12	10	1.5	0.5	2660	2574
9	12	10	1.0	1.0	2650	2652
10	12	10	1.75	0.25	2627	2687
11	14	10	1.5	0.6	2685	2766
12	12	15	1.5	0.4	2600	2535
13	14	10	--- VOCODED BUT UNQUANTIZED ---			
14	---------- ORIGINAL WAVEFORM, DIGITIZED AND RECONSTITUTED					

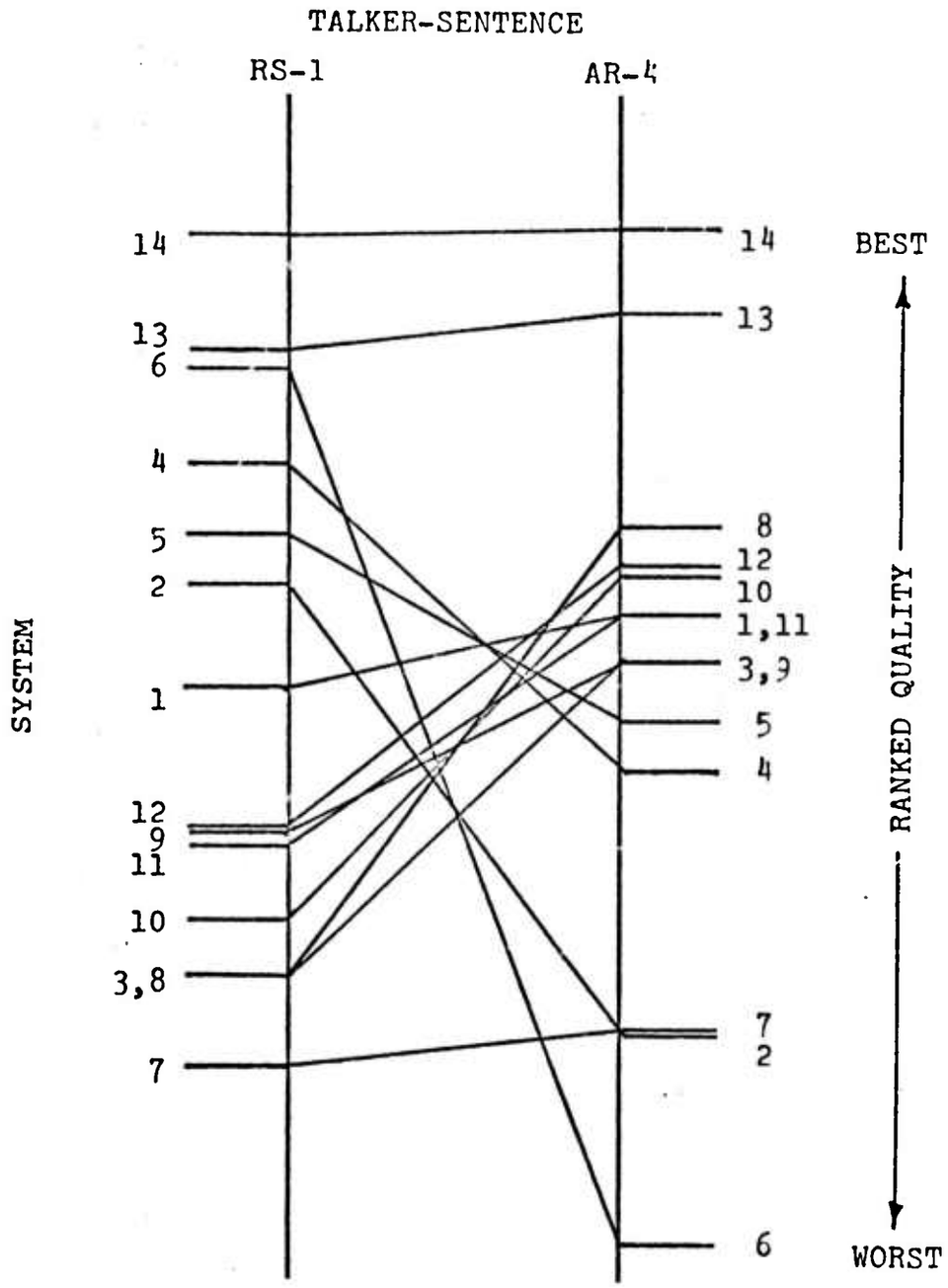

Fig. 5. Projections of the points representing systems shown in Fig. 4 onto the vectors representing talker-sentence combinations RS-1 and AR-4.

occasional /w/ or /r/) are stops or affricates, which are characterized by very abrupt changes of spectrum and of envelope. Talker RS was a female with a moderate speaking rate, whereas talker AR, also a female, was the fastest speaker in the sample. System 6, which performed very well on sentence RS-1, was a 10-pole system with a large frame size of 25 msec, and a small quantization step size of 0.2 dB. These parameters combine to give adequate coding of a slowly changing spectrum, as in sentence 1, but quite inadequate coding of sentence 4, with its repeated abrupt changes.

In contrast to system 6, system 8 was judged to do a better job with AR-4 than with RS-1. System 8 had 12 poles, a frame size of 10 msec, a quantization step size of 0.5 dB, and was transmitted at a variable rate. Again, the result is understandable in terms of the properties of the sentences and the speech characteristics of the talkers.

The fact that some of the systems used variable-rate transmission (systems 8-12) also influenced the bit rates achieved in the different sentences. The variable-rate systems were chosen so that their average bit rate was about 2600 bits/sec, equal to that of the fixed rate systems. When presented with a very easy sentence, such as sentence RS-1, the bit rate of system 8 (a variable-rate system) dropped to 2306 bits/sec. On sentence AR-4, on the other hand, the bit rate rose to 2921. This factor clearly played a part in influencing the ralative quality of fixed- and variable-rate systems, as one would expect.

Figure 6 shows the effect of the sentence material on the preference judgments in another way. Note that the relative positions of systems 4, 5 and 6 on the quality scale decrease as

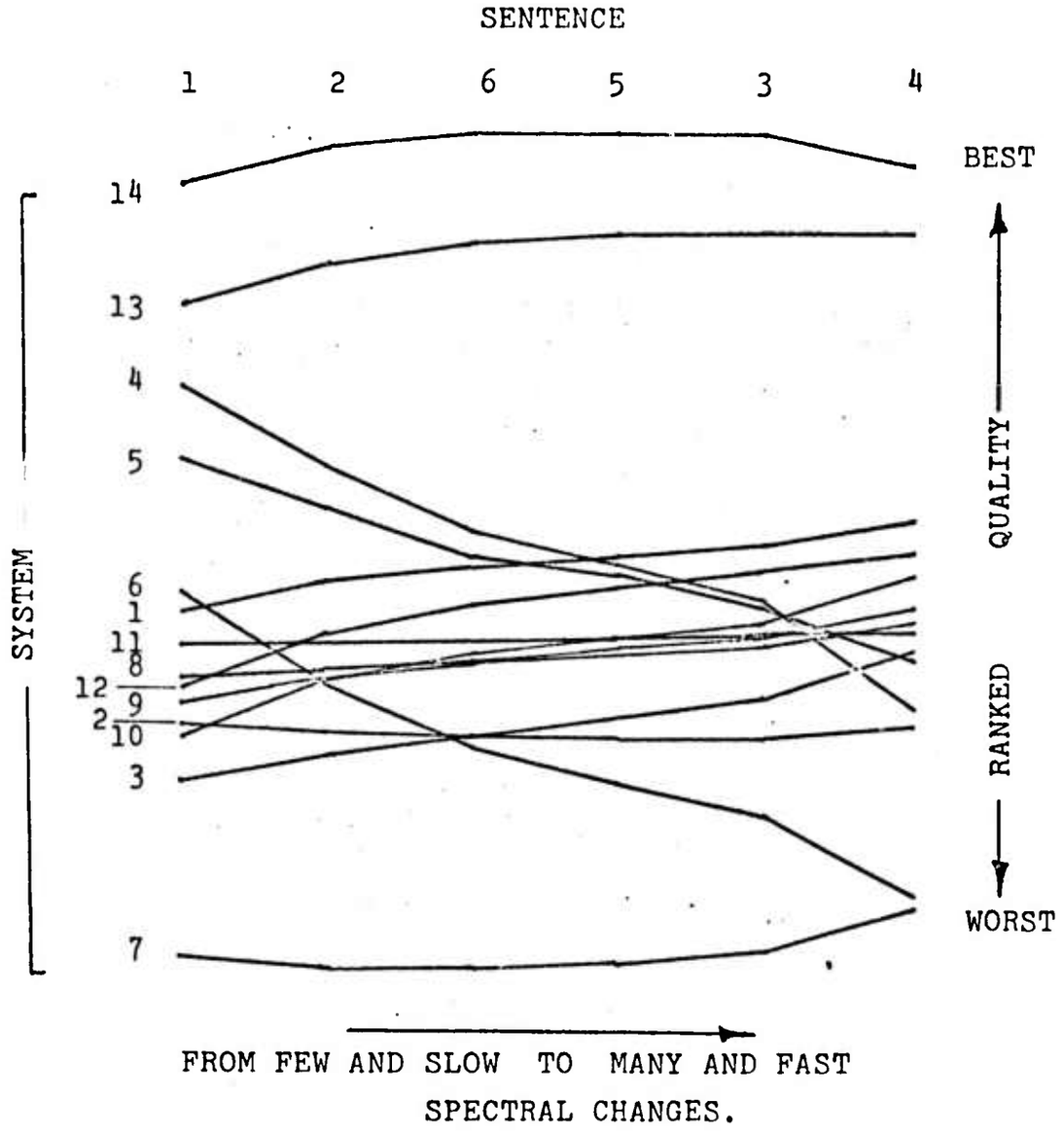

Fig. 6. Effects of sentence materials on relative preferences for tested systems (see Tables 3 and 4).

the sentences become less and less smooth. Conversely, systems 1, 3, 10 and 12 become progressively better under the same condition. (For the parameters of these systems, see Table 4.)

Figure 7 shows a similar representation of the effect of the talker as a determinant of preference. The talkers are arranged in the order of the fundamental frequencies of their voices (high to low, left to right). It is of interest to note that two of the systems, 2 and 6, seemed to be particularly sensitive to this aspect. Both did relatively well with high-pitched voices, and very poorly with low. These two systems, together with system 7, were the only ones that used a 10-pole spectral approximation, all others using either 12 or 14 poles. It is well known that low-pitched voices require more poles for adequate matching of their spectra, so the poor performance of systems 2, 6, and 7 on low-pitched voices is again easily explainable. We should note that the results shown in these figures do not establish whether the outputs of the systems vary in absolute quality as functions of sentence material or talker characteristics, only that they change relative to each other.

Few studies have focused on the role of the nature of the speech material or the characteristics of the talker's speech as determinants of the outcomes of quality evaluations. Those that have, however, have produced results that agree with our results in showing that sentence and talker effects can be substantial, and, if not taken into account, can lead to faulty interpretations of evaluation data. Some evidence of the importance of proper selection of test sentences, for example, is presented by Pachl, Urbanek, and Rothauser (1971). In their study, the percentage of judgments favoring a given system over others in a direct comparison task varied greatly depending on the sentence that was used for the comparison. These investigators

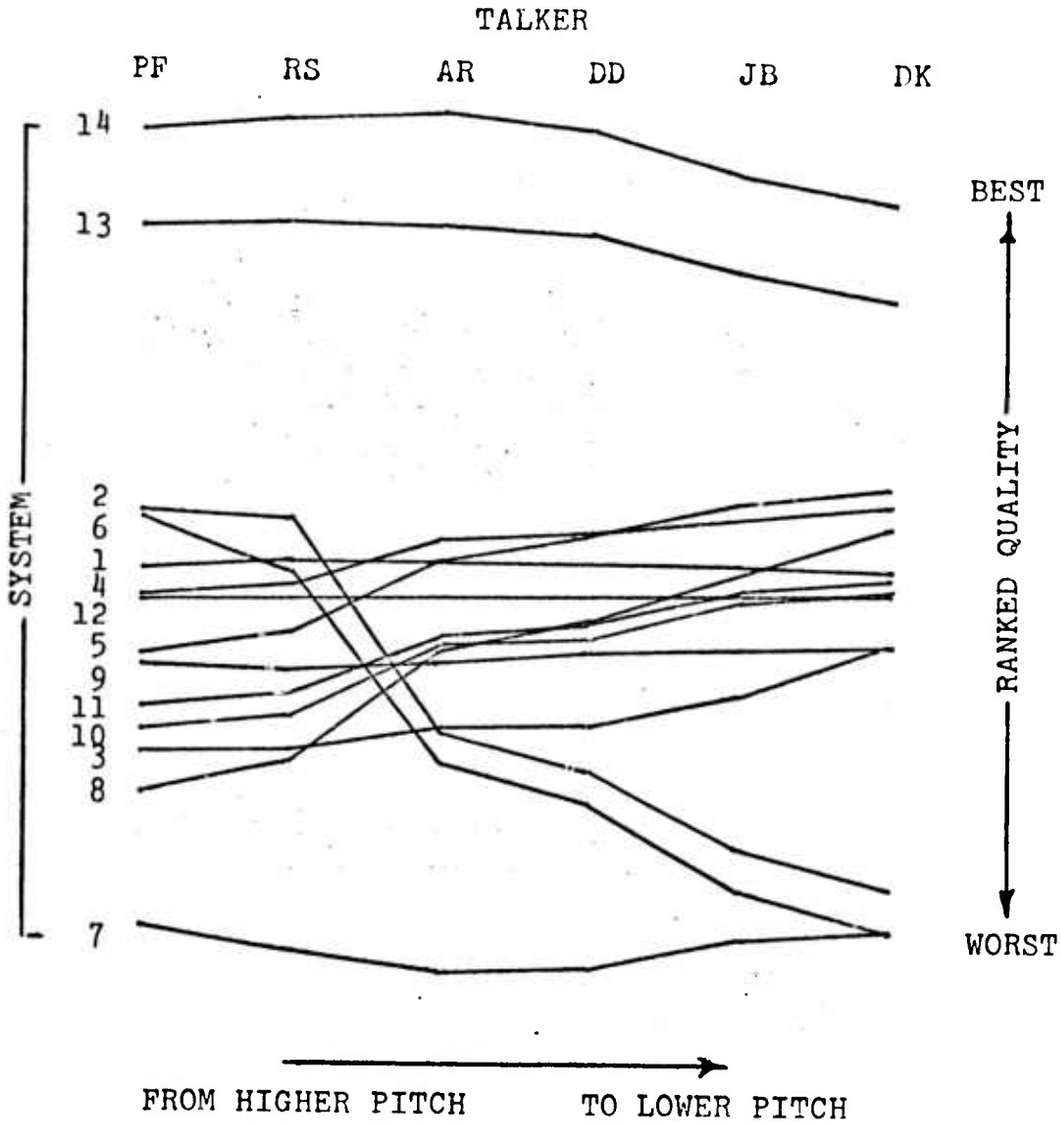

Fig. 7. Effects of talker characteristics on relative preferences for tested systems (see Tables 2 and 4).

concluded from this finding that, if meaningful results are to be obtained from preference judgments, the same test materials must be used with all systems. We agree with this point, but suggest that invariance of materials across systems is, by itself, an insufficient requirement. It is also essential that the material that is used with a given system be as broadly representative of the vagaries of speech as is practically feasible, and that the same broad sampling of materials be used with every system. Use of material that is invariant across systems, but not sufficiently representative of speech in general, could yield highly misleading results by producing a rank ordering of systems that would hold only for speech with the particular characteristics of the sample.

There is a second, and equally important, reason for using carefully selected materials, representing a broad range of characteristics. Such an approach provides the opportunity for acquiring information concerning the strengths and weaknesses of individual systems to deal with specific aspects of speech or voice characteristics. A clear example of this advantage was given above, in discussing the relative strengths and weaknesses of variable-rate systems as against fixed-rate systems.

3. JUDGING INDIVIDUAL SPEECH QUALITIES

An approach to speech evaluation quite different from that of obtaining judgments of its overall quality is that of trying to assess it with respect to specific aspects or features. One may ask a listener to attend to one or more specific aspects of an utterance, such as its loudness (Coolidge & Reir, 1959), its degree of nasality (Stevens, Nickerson, Rollins, & Boothroyd, 1975), the appropriateness of its timing or rhythm (Boothroyd, Nickerson, & Stevens, 1975).

A commonly used method for obtaining multidimensional descriptions of complex stimuli is that of semantic differential scaling (Osgood, 1952). The approach is illustrated by an experiment by Kerrick, Nagel, and Bennett (1969), one point of which was to determine the extent to which the concepts of loudness and noisiness could be operationally distinguished (see Table 5). Loudness and noisiness proved to be nearly equivalent descriptors in this study, the correlation between the ratings on these dimensions being .96. Neither loudness nor noisiness correlated highly, however, with acceptability (.30 and .43, respectively). A 2-dimensional plot of the noisiness of the stimuli against their acceptableness suggested that the acceptability of a given level of perceived noisiness depends on the nature of the sound; higher levels of noisiness are acceptable for musical sounds than for vehicle sounds, and for vehicle sounds than for "artificial" sounds.

Another suggestive result from this study came from a comparison of the reactions of two listeners to the same sound (broad-band noise). Subjects were not told the source of the sounds, but were asked to identify them. One subject identified this sound as "air blowing" and another as a jet flyover. The former subject judged the sound to be louder and noisier, but more acceptable than did the latter, suggesting that the degree of acceptability of a given level of perceived noisiness may depend not only on the nature of the sound but of its assumed origin. While these results were obtained with non-speech stimuli, they point out the importance of variables other than stimulus properties per se as determinants of individual preferences, and it seems likely that similar effects might be found with speech.

The approach of comparing speech with respect to specific characteristics has been criticized on the grounds that it is not

Table 5. Scaling dimensions used by Kerrick, Nagel, and Bennett (1969) for semantic-differential description of sounds. Listeners rated each sound with respect to each of these dimensions on a 7-point scale.

good	---	bad
far	---	near
unfamiliar	---	familiar
noisy	---	quiet
fast	---	slow
smooth	---	rough
natural	---	unnatural
soft	---	loud
passive	---	active
acceptable	---	unacceptable
high	---	low
delicate	---	rugged
pleasant	---	unpleasant
narrow	---	wide
light	---	heavy

clear how to derive a measure of overall quality from the results of such comparisons (Rothauser, Urbanek, & Pachl, 1968; Tedford & Frazier, 1966). To the extent that one is interested in differences with respect to specific features per se, as opposed to differences in overall quality, this limitation is irrelevant. But if the primary interest is in overall quality differences, it clearly carries some weight.

Another difficulty that has been pointed out by Rothauser, Urbanek, and Pachl (1968) is that a given qualitative descriptor can mean different things to different listeners or in different contexts. "Naturalness," for example, might be used to represent the degree to which a transmission preserves the voice characteristics of the speaker, when judging telephone circuits; when used in connection with synthetic speech, however, it might represent the degree to which the speech sounds human.

Notwithstanding these difficulties, it seems to us worthwhile to explore the possibility of having listeners characterize vocoded speech in terms of specific perceptual properties. As a step in this direction, a listening test was conducted for the purpose of identifying descriptive terms that listeners who are not familiar with vocoding techniques would consider appropriate, or useful, for characterizing vocoded speech.

The test was carried out in two stages. In the first stage, the listeners were requested to list adjectives or phrases that they considered descriptive of the speech to which they were listening. In the second stage, they were provided with lists of words and phrases, and asked to judge the appropriateness of each of the items on the lists to the speech.

The speech samples were generated by two talkers, an adult male and an adult female (JB and AR of Table 3), speaking each of two different sentences (sentences 5 and 6 of Table 1). Nine versions of each sentence spoken by each talker were used in the test. One version was a digitized, but otherwise unprocessed, sample (system 14 of Table 4). The other eight versions were produced by the LPC vocoder with the parameter values shown in Table 6. Table 6 also gives transmission bit rates for each of the samples, which range between 1060 and 3159 bits/sec. The considerable range of bit rates was intended to produce speech samples that would vary significantly in quality. It was hoped that the orthogonal variation of the system parameters would produce perceptible differences even within a quality class.

Sentences were heard in pairs. The first member of a pair was always the unprocessed version of the sentence; the second member was one of the eight processed versions of the same sentence spoken by the same talker. The reason for this arrangement was to provide listeners with a continual reminder of what the speech they were hearing would sound like when not processed, so they could judge the effects of the vocoding the more accurately. Listeners were encouraged to attend to the ways in which the standard (unprocessed) and test (processed) samples differed.

Listeners were 17 undergraduates, 2 females and 15 males, at Tufts University, members of a class on Human Factors Engineering in the School of Engineering Design. They participated in this test as part of an optional class exercise.

After the instructions were read, the tape recording was started and the listeners were requested to listen to several items and then begin making their list of descriptors. After 10 minutes,

Table 6. System specifications and transmission rates for material used in attempt to identify descriptors of perceptual properties of vocoded speech.

System #	Specifications			Bit Rate				System Average Bit Rates
	Predictor Order	Frame Rate	Quantization Step Size (dB)	Male (JB) Sentence #1	Male (JB) Sentence #2	Female (RS) Sentence #1	Female (RS) Sentence #2	
1	9	Variable	2.0	1180	1115	1120	1060	1119
2	12	Variable	2.0	1438	1448	1466	1431	1446
3	9	Variable	0.5	1552	1462	1447	1364	1456
4	9	Constant	2.0	1770	1745	1795	1775	1771
5	12	Variable	0.5	1821	1844	1851	1810	1831
6	12	Constant	2.0	2354	2325	2382	2359	2355
7	9	Constant	0.5	2520	2495	2545	2525	2521
8	12	Constant	0.5	3154	3135	3132	3159	3157

Sentence #1. The little blankets lay around on the floor. (Sentence 5 of Table 1)

Sentence #2. The trouble with swimming is that you can drown. (Sentence 6 of Table 1)

these lists were gathered and previously prepared check lists were distributed. The listeners continued the test (stage 2) by rating each of the words and phrases on these lists as to its appropriateness as a descriptor of the processed speech they were hearing. Ratings were made on a scale ranging from 0 (completely inappropriate) to 9 (perfect descriptor). During the time the listeners were working on the prepared lists, another list was composed consisting of items produced by the listeners during the first stage that were not on the prepared lists. This new list was put on the blackboard, and the listeners continued the test by assigning scale values to these new terms.

The check lists that were prepared before the listening test contained 127 words and 26 phrases (see Tables 7 and 8) that had been selected by the investigators as being possible candidates as descriptors of vocoded speech.

Table 9 shows the words and Table 10 the phrases produced by the listeners during stage 1 of the test. Table 11 shows the ten words receiving the highest ratings, considering all of the listeners, and also considering two subsets of "best" listeners. "Best" was defined somewhat arbitrarily in terms of the degree to which the scale values assigned by a listener to individual words deviated from the means of the scale values assigned by the entire group.

Table 7. Prepared list of words whose appropriateness to vocoded speech quality was rated by 17 subjects in Task 2.

In some of the pairs, the second sentence has a _____quality.

1___blary
2___boomy
3___bouncy
4___brassy
5___breathy
6___burbly
7___buzzy
8___chirpy
9___choppy
10___chattery
11___clean
12___clicky
13___clipped
14___coarse
15___computer-like
16___crackly
17___creaky
18___crisp
19___croaky
20___damped
21___dead
22___deep
23___diffused
24___disconnected
25___distinct
26___distorted
27___drone-like
28___dull
29___eddying
30___electronic
31___even
32___frizzy
33___flat
34___fluctuating
35___fuzzy
36___garbled
37___grating
38___grinding
39___gruff
40___gurgly
41___guttural
42___hissy
43___hollow
44___human
45___hum-like
46___hushed
47___husky
48___indistinct
49___jangling
50___jerky
51___mellow
52___metally
53___monotone
54___murmury
55___musical
56___muted
57___nasal
58___natural
59___noisy
60___oscillating
61___piercing
62___hi-pitched
63___pulsating
64___pure
65___raspy
66___reed-like
67___regular
68___resonant
69___reverberant
70___rich
71___ringy
72___rough
73___scratchy
74___sharp
75___sharp-edged
76___shivery
77___shrill
78___silvery
79___slurred
80___smooth
81___smooth-edged
82___soft
83___spitty
84___spluttery
85___sputtery
86___squawky
87___squeaky
88___steady
89___stifled
90___strained
91___strident
92___subdued
93___telephonic
94___throbbing
95___tinny
96___trill
97___twangy
98___tweeting
99___twittery
100___unbroken
101___unclean
102___undulatory
103___uneven
104___vibrant
105___warbly
106___wavery
107___wheezy
108___whirring
109___whispery
110___wobbling
111___yodelling
112___whistling
113___tinkling
114___thin
115___swishing
116___screeching
117___rumbling
118___rippling
119___radio-static
120___quavering
121___harsh
122___full
123___fluttering
124___flat
125___echoing
126___clear
127___broken

Table 8. Prepared List of phrases whose appropriateness to vocoded speech quality was rated by 17 subjects in Task 2.

In some of the pairs, the second sentence sounds as if _____.

1____the speaker is feeble.
2____the speaker is hoarse.
3____the speaker is being interrupted.
4____the speaker is muffled.
5____the speaker is being shaken.
6____the speaker is straining to speak.
7____the speaker is trembling.
8____the speaker is under water.
9____the speaker has a head cold.
10____the speaker is speaking through a barrel.
11____the speaker is normal.
12____the speaker is talking in a cave.
13____the speaker is talking in a high wind.
14____the speaker is non human.
15____the speaker is having difficulty breathing.
16____the speaker is old.
17____the speaker is tired.
18____the speaker is speaking through a pipe.
19____the speaker's pitch is changing.
20____the speaker has something in his/her mouth.
21____the speaker sounds like a parrot talking.
22____the speaker has a lisp.
23____the speaker is a machine.
24____the speaker is nervous.
25____the speaker is talking while being vibrated.
26____the speaker is talking over a telephone.

Table 9. Words and phrases produced by listeners in Task 1.

		Frequency of Appearance			Frequency of Appearance
1.	accented	1	28.	monotone*	2
2.	blanketed	1	29.	more distant	
3.	blurred	1	30.	more echoic	1
4.	bored	2	31.	muffled*	11
5.	bubbling	1	32.	narcious	1
6.	chilly	1	33.	nasal*	10
7.	choking	1	34.	n n n-ing the n's	1
8.	congested	1	35.	noisy*	2
9.	const icted	1	36.	normal*	1
10.	cotton	1	37.	scratchy	1
11.	diminished clarity	2	38.	sick	1
12.	distorted*	4	39.	singing	1
13.	dry	1	40.	slurring*	2
14.	dull*	1	41.	soft*	2
15.	fading	1	42.	sorrowful	1
16.	flat*	1	43.	speeded up	1
17.	fuzzy*	1	44.	squeakier*	1
18.	garbled*	2	45.	sshhing together	1
19.	harsher	1	46.	static	1
20.	hesitation	1	47.	stuffed up	2
21.	hollow*	2	48.	sustaining	1
22.	humming	1	49.	tight-lipped	1
23.	hurried	1	50.	unenthusiastic	1
24.	less distinguishable	1	51.	vibrating	1
25.	level	1	52.	whinning	1
26.	lispy	1	53.	whispering*	1
27.	messed up	1	54.	wobbly*	1
			55.	wonder	1

*Words which also appeared on our prepared lists of sound quality descriptors, Table 7 and 8.

Table 10. Phrases produced by listeners in Task 1.

1. telling a story
2. the lady states a sad fact
3. high and low peak quality
4. r's accentuated
5. vary-accent
6. in the water/underwater*/speaking underwater
7. put a weight on it
8. not opening his mouth
9. metallic edge of the voice
10. nasal drawl
11. high frequency enhancement
12. less bored
13. has congested sinuses
14. has a cold*
15. pinching his nose/holding his nose/has a clothes pin on his nose
16. loss of volume
17. slurring syllables
18. damaged vocal cords
19. clipped words
20. speaking into something (a container)
21. someone has put their hand partially over the speaker's mouth/has hand over face
22. smothered in cotton
23. inside a bell
24. speaker is in a chamber
25. cleft palate
26. out of breath
27. speaking through a stretched membrane or through a balloon
28. has post nasal drip
29. from New England
30. running words together

---

* Phrases which also appeared on our prepared lists of sound quality descriptors, Tables 7 and 8.

Table 11. "Top ten" words.

All	Best 9	Best 3
nasal	nasal	nasal
muffled	muffled	muffled
distorted	distorted	fuzzy
monotone	head cold	distorted
blanketed	garbled	stuffed up
fuzzy	dull	muted
head cold	monotone	blanketed
dull	blanketed	head cold
garbled	fuzzy	damped
muted	slurred	parrot-like

4. PLANS FOR THE FUTURE

1. Testing of the listening procedure described in Section 7.3 of the preceding QPR. Stimulus materials have been prepared for this test and we plan to run subjects on it during the next quarter.

2. Conducting of phoneme-specific tests described in Section 7.5 of the preceding QPR.

3. Further investigation of the suitability of multidimensional scaling techniques for speech-quality evaluation.

4. Extension of the work described in Section 3 of this report. This will probably include the development of a set of semantic-differential scales for use in further speech-quality evaluation studies. These scales will be chosen on the basis of the results obtained in the study described in Section 3.

5. Preparation of additional recorded speech samples to be used for processing by several of the LPC vocoder systems currently under development. We will use the sentence materials that we have developed and also, probably, the talkers for whom we have made speech-characteristic measurements. The plan is to make the recordings under less than ideal conditions, as was agreed upon at the contractor meeting at BBN in June. The Lincoln Laboratory group will participate with us in this activity.

6. Establishment of a library of vocoded speech materials to be used in quality-evaluation studies. These materials will be supplied by the various contractors. The purposes for

establishing this library are to make it possible to make direct comparisons among specific systems of interest, and to study the effects of particular vocoding techniques or parameter settings on specific qualitative aspects of the resulting speech.

References

Boothroyd, A., Nickerson, R. S., & Stevens, K. N. Temporal patterns in the speech of the deaf--An experiment in remedial training. Journal of Speech and Hearing Disorders, in press.

Carroll, J. D. Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. Nerlove (Eds.), Multidimensional scaling: Theory and applications in the behavioral sciences. Vol. 1. Theory. Nes York: cademic Press, 1972.

Carroll, J. D. & Wish, M. Multidimensional perceptual models and measurement methods. In E. C. Carterette & M. P. Friedman (Eds.), Handbook of perception. Vol. 2. New York: Academ Press, 1974.

Coolidge, O. H. & Reir, G. C. An appraisal of received telephone speech volume. Bell System Technical Journal, 1959, 38, 877.

Kerrick, J. S., Nagel, D. C., & Bennett, R. L. Multiple Ratings of sound stimuli. Journal of the Acoustical Society of America, 1969, 4, 1014-1017.

Lieberman, P. Same acoustic measures of the fundamental periodicity of normal and pathological larnyges. Journal of the Society of America, 1963, 35, 344-353.

McDermott, B. J. Multidimensional analyses of circuit quality judgments. Journal of the Acoustical Society of America, 1969, 45, 774-781.

McGee, V. E. Semantic components of the quality of processed speech. Journal of Speech and Hearing Research, 1964, 7, 310-323.

McGee, V. E. Determining perceptual spaces for the quality of filtered speech. Journal of Speech and Hearing Research, 1965, 8, 23-28.

Osgood, C. E. The nature and measurement of meaning. Psychological Bulletin, 1952, 49, 197-237.

Pachl, W. P., Urbanek, G. E., & Rothauser, E. H. Preference evaluation of a large set of vocoded speech signals. IEEE Transactions on Audio and Electroacoustics, 1971, AU-19, 216-224.

Rothauser, E. H., Urbanek, G. E., & Pachl, W. P. Isopreference method for speech evaluation. Journal of the Acoustical Society of America, 1968, 44, 408-418.

Stevens, K. N., Nickerson, R. S., Rollins, A., & Boothroyd, A. Use of a visual display of nasalization to facilitate training of velar control for deaf speakers. Bolt Beranek and Newman Inc. Report No. 2899, September, 1974.

Tedford, W. H., Jr. & Frazier, T. V. Further study of the isopreference methods of circuit evaluation. Journal of the Acoustical Society of America, 1966, 39, 645-649.