LOCALIZATION OF SOUND

Part 5. Auditory Perception

by

United Research, Inc.
Cambridge, Massachusetts

ABSTRACT. The delay theory of auditory localization
is presented which provides a model for attention,
location, and recognition wherein the function of the
basilar membrane is that of a tapped delay line. It is
shown that the same theory may be applied to speech
recognition phenomena to show a consistency of the
auditory mechanism in the perception and utilization
of acoustic information. A new theory of information
handling in the nervous system based on photon
emission is described.

The design and construction of insertion type
electrostatic headphones is discussed and perform-
ance data given. Localization tests were conducted
using the headphones which show that effective aural
coupling is achieved.

Improvements in devices developed for use in
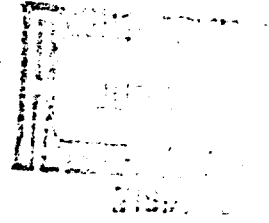porpoise communication research are presented.

U.S. NAVAL ORDNANCE TEST STATION

China Lake, California

January 1965

FOREWORD

The purpose of this report is to describe studies in the delay theory of auditory localization which provides a model for attention, location, and recognition wherein the function of the basilar membrane is that of a tapped delay line.

This work was conducted by United Research, Inc., Cambridge, Mass., under Contract No. N123-(60530) 35401A issued by the U.S. Naval Ordnance Test Station. The information herein covers work that completed this contract in October 1964.

This is Part 5 of a series of reports that were issued covering various aspects of the subject. The titles of Parts 1, 2, 3, and 4 are as follows:

> Part 1. Characteristics of Human Localization of Sound (December 1962)
>
> Part 2. The Mechanism of Human Localization of Sounds With Applications in Remote Environments (December 1962)
>
> Part 3. A New Theory of Human Audition (December 1963)
>
> Part 4. Further Developments in Sound Localization (April 1964)

The text pages in this report were reproduced in facsimile from a report issued by the contractor.

WM. B. MCLEAN
Technical Director

INITIAL DISTRIBUTION

6 Chief, Bureau of Naval Weapons
    DLI-31 (2)
    RRRE (1)
    RU (1)
    RUTO (1)
    RUTO-2 (1)
3 Special Projects
    SP 20 (1)
    SP 22 (1)
    SP 27 (1)
1 Chief, Bureau of Ships (Code 560)
1 Chief of Naval Operations
2 Chief of Naval Research
    Code 104 (1)
    Code 466 (1)
1 David W. Taylor Model Basin
1 Fleet Anti-Air Warfare Training Center, San Diego (Guided Missile Section)
1 Naval Advanced Undersea Weapons School, Key West
1 Naval Air Development Center, Johnsville
1 Naval Engineering Experiment Station, Annapolis
1 Naval Ordnance Laboratory, Corona
1 Naval Ordnance Laboratory, White Oak
1 Naval Postgraduate School, Monterey (Library, Technical Reports Section)
1 Naval Research Laboratory
1 Naval Torpedo Station, Keyport (Quality Evaluation Laboratory, Technical Library)
1 Naval Underwater Ordnance Station, Newport
1 Naval Underwater Weapons Systems Engineering Center, Newport
2 Navy Electronics Laboratory, San Diego
1 Navy Mine Defense Laboratory, Panama City
1 Navy Underwater Sound Laboratory, Fort Trumbull
1 Navy Underwater Sound Reference Laboratory, Orlando
1 Norfolk Naval Shipyard (Underwater Explosion Research Division)
1 Pacific Missile Range, Point Mugu
1 Submarine Development Group 2
1 Army Research Office, Arlington (Lester H. Geiger)
1 Atlantic Missile Range
20 Defense Documentation Center (TISIA-1)
1 National Aeronautics & Space Administration

CONTENTS

ACKNOWLEDGMENT

CHAPTER 1

INTRODUCTION

1.1   This report describes the latest results of research in auditory localization which began in 1960.

Three significant results are reported. One, the extension of the original hypothesis regarding time delays in auditory localization to a theory of perception supported by mathematical models and consistent with observations and experiments. Second, the design and construction of satisfactory electrostatic headphones which proved to be the most difficult component to develop in a system to reproduce localization information accurately. Third, the application of certain theoretical results to speech recognition and the subsequent development of devices to establish a basic man-to-porpoise communication link.

The theory reported here is based on the fundamental concept that we derive knowledge of our environment by mentally inverting a transformation introduced on the observed space by the mechanism of perception. The form of the transformation ascribed to human audition, speech, color vision, etc. is that of time delays. The model described for inverting such a transformation is realizable in the human nervous system. However, to do so, different functions must be assigned to certain elements of perception mechanisms. For example, in hearing, the basilar membrane is no longer considered to function as a resonant structure, but rather as a delay line from which the inverse transform computation is made. Such radical departure from the popular concepts implies a rethinking of the traditions by which new theories are judged and evaluated. It has been our experience that the theory presented here, while providing an understanding of many observations, is marked with the tag of controversy. Nevertheless, while traditional thought holds sway, developments continue to indicate the correctness of our

thought. Recently, it was reported that intelligible speech may be trans-
mitted on a 10-cps bandwidth. This accomplishment is achieved by a de-
vice that is said to be an electronic representation of the human ear. The
cochlea and basilar membrane are simulated by a delay line with detectors
tapped along its length. This is the function of the basilar membrane in the
theory proposed.

1.2   A system for transferring localization information was improved
by refining the design and construction of electrostatic headphones re-
ported earlier. The development was created by the lack of availability of
headphones of requisite bandwidth. While the design reported here has
proved effective in accurate transfer of localization information, further im-
provements can be made.

1.3   The work in porpoise communication (ref. 3.4) was continued with
emphasis on equipment improvements and the development of a meta-language.

CHAPTER 2

THEORY

by

Dwight W. Batteau

2.1 Summary

During this contract, theoretical aspects of the derived attention functions for human hearing were examined and hypotheses were formed concerning means of realizing such functions in the human nervous system. The transformations applicable to the role of the pinna in localization, speech formation and recognition, room reverberation, and object identification, as performed by the dolphin, were examined. Estimates of improvement in signal selection by attention and parameters in recognition were examined.

2.2 Pinnae Attention

The term "cocktail party effect" has been applied to the ability of a man to pay attention selectively to a desired conversation (or other sound) in the presence of other conversations or sounds.

An expression can be written for the function of the pinna, equation (2.2.1)

$$H(s) = P(s) \sum_{n=0}^{N} a_n \epsilon^{-s\tau_n} \qquad (2.2.1)$$

$H(s)$ = sound at eardrum

$P(s)$ = sound reaching the pinna

a_n = coefficient of reflection for the nth delay path

τ_n = delay in the nth delay path

9

It is convenient to normalize equation (2.2.1) by assigning the following values:

$$a_o = 1$$

$$\tau_o = 0$$

An attention function can be constructed applicable to equation (2.2.1) as follows in equation (2.2.2).

$$A(s) = \sum_{n=0}^{M} b_n \epsilon^{-s(\tau_M - \tau_n)} \qquad (2.2.2)$$

This function is constructed by reversing the ordering of the delays in equation (2.2.1). If (2.2.2) is applied to equation (2.2.1) the following result is obtained, equation (2.2.3).

$$H(s)\,A(s) = P(s)\epsilon^{-s\tau_M}\left[\sum_{n=0}^{N} a_n b_n + \sum_{j=0}^{N}\sum_{\substack{k=0 \\ j \neq k}}^{N} a_j b_k \epsilon^{-s\tau_j} \epsilon^{+s\tau_k}\right] \qquad (2.2.3)$$

Diagrammatically, if P(s) is assumed to be a pulse, equation (2.2.1) produces the result shown in Figure 2-1 for N = 3.
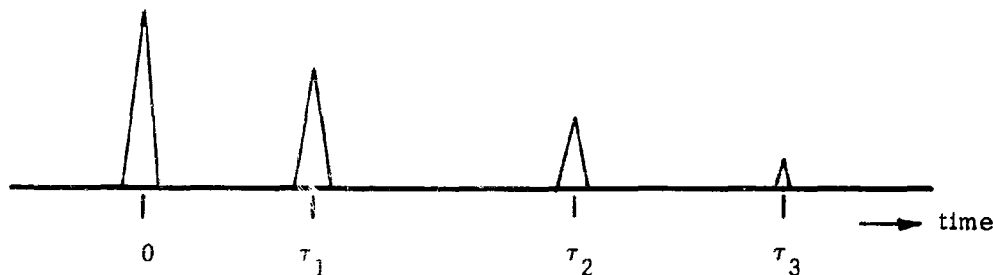
Figure 2-1. The pinna transform of a pulse, assuming four paths.

Applying equation (2.2.3) to the same situation results in the signal sketched in Figure 2-2.
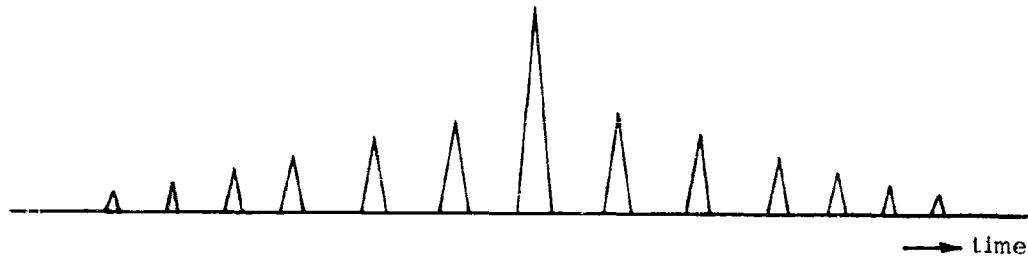
Figure 2-2. The result of applying equation (2.2.3).

The central pulse in Figure 2-2 results from the coherent term in equation (2.2.3), as given below.

$$P(s) \sum_{n=0}^{N} a_n b_n \tag{2.2.4}$$

The pulses on either side of the central pulse in Figure 2-2 results from the cross terms of equation (2.2.3).

$$P(s) \sum_{j=0}^{N} \sum_{k=0}^{N} a_j b_k \epsilon^{-s\tau_j} \epsilon^{+s\tau_k} \tag{2.2.5}$$
$$j \neq k$$

If a second sound source $Q(s)$ is located at a separate place from $P(s)$, the transformation is the same form, but has different reflection coefficients and delays. This is presented as equation (2.2.6).

$$J(s) = Q(s) \sum_{k=0}^{N} c_k \epsilon^{-s\tau_k} \tag{2.2.6}$$

If equation (2.2.2) is applied to (2.2.6) the result is equation (2.2.7).

$$J(s) A(s) = Q(s) \epsilon^{-s\tau_N} \sum_{j=0}^{N} \sum_{k=0}^{N} b_j c_k \epsilon^{-s\tau_j} \epsilon^{+s\tau_k} \qquad (2.2.7)$$

The result of equation (2.2.7) if $Q(s)$ is a pulse can be sketched in Figure 2-3.
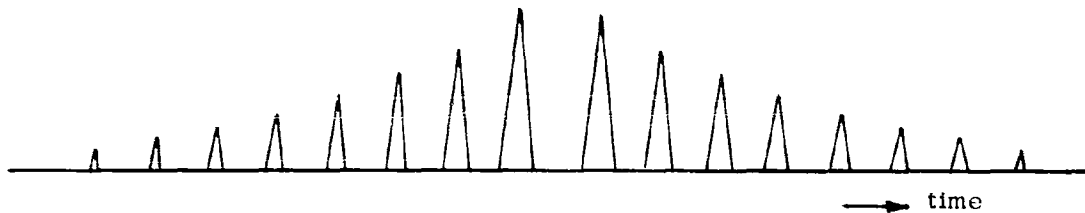
Figure 2-3. The result of applying equation (2.2.2) to equation (2.2.6).

The significant difference between the two results is the large central spike produced from the reverberant pulse train when the transformations are matched.

Let us assume that the two signals $P(s)$ and $Q(s)$ are now independent sources of white noise. By definition

$$\int_{-\infty}^{\infty} P(t) \, P(t + \tau) \, dt = 0 \qquad (2.2.8)$$
$$\tau > 0$$

$$\int_{-\infty}^{\infty} Q(t) \, Q(t + \tau) \, dt = 0 \qquad (2.2.9)$$
$$\tau > 0$$

$$\int_{-\infty}^{\infty} Q(t) \, P(t) \, dt = 0 \qquad (2.2.10)$$

Thus the delayed signals represent independent signal powers because of the zero value of cross correlations.

In order to compute the effect of signal processing of the kind described, we can make initially two assumptions:

(1) All the coefficients are unity.

(2) The addition is equivalent to adding voltages in an electrical signal.

If there are N terms in the reflection system, then there is a total of N^2 terms in the correlation output. In the case of the coherent term, N of these are added. Thus the powers resultant can be computed. For the monaural or single channel, the power due to P(s) is as follows:

$$_1U_p = N^2 + N^2 - N \qquad (2.2.11)$$

$$_1U_p \overset{\Delta}{=} \quad \text{power due to P(s) after the attention transformation in a single channel}$$

For the same case, the power due to Q(s) is as follows:

$$_1U_q = N^2 \qquad (2.2.12)$$

$$_1U_q \overset{\Delta}{=} \quad \text{power due to Q(s) after the attention to P(s) transformation in a single channel}$$

The resultant power ratios are expressed in equation (2.2.13)

$$\frac{_1U_p}{_1U_q} = {_1R_{p,q}} = 2 - \frac{1}{N} \qquad (2.2.13)$$

$$_1R_{p,q} \overset{\Delta}{=} \quad \text{the ratio of powers in a single channel due to two signals, one of which (p) } \text{undergoes an attention transformation}$$

In the binaural case, or for two channels the resultant powers due to attention to one of the signals is as follows:

7

$$2U_p = 4N^2 + 2N^2 - 2N \qquad\qquad (2.2.14)$$

$2U_p \overset{\Delta}{=}$ power due to P(s) after attention transformations in two channels

$$2U_q = 2N^2 \qquad\qquad (2.2.15)$$

$2U_q =$ power due to Q(s) after attention transformations in two channels

The resultant power ratios are expressed in equation (2.2.16).

$$\frac{2U_p}{2U_q} = 2R_{p,q} = 3 - \frac{1}{N} \qquad\qquad (2.2.16)$$

The limits for the two cases are 2 and 3 respectively, or 3 db and 4.8 db. This indicates the limits of selection corresponding to differences in power of two separated sound sources, differently transformed by the pinnae or the environment.

If a comparison is made for two channel selection on time difference alone in an anechoic environment, the two channel ratio of powers is 2 or a limit of 3 db. Thus reverberation can be used to improve selection in hearing. Both the reverberation due to the pinnae and that due to the environment can be used.

2.3   Function Construction

The ideal detector is one which does not alter the characteristics of the signal. An ideal microphone, for example, would have no resonances and produce no reflections. If we assume that the Organ of Corti is the sonic detector for hearing, it should approach these requirements in view of its excellent performance. In this view, the cochlea provides a model of acoustic termination. If the system is viewed as a straight element, it shows the taper of acoustic terminations (as in anechoic chambers) and

8

the continuous change in apparent impedance towards the small end.
This is sketched in Figure 2-4.
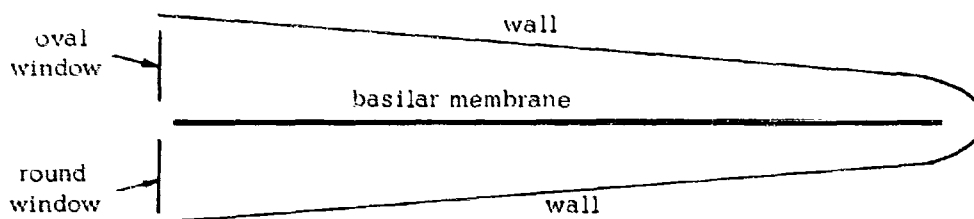
Figure 2-4. The cochlea as an acoustic termination.

While the high frequencies could be terminated in this manner, the
scale is small for low frequencies. However, the flexible central membrane
and the round window (as a pressure release orifice) can provide anechoic
termination for the low frequencies. The arrangement of dense and spongy
bone surrounding the cochlea also provides a model of acoustic isolation.
From an engineering viewpoint, the detector is isolated, protected and
terminated anechoically. From this viewpoint, there is no point in
examining mechanical resonance as a model of tone or pitch detection.
However, the anechoic model provides an ideal delay line, distributing
the signal over the nerve endings on the basilar membrane. If we consider
20 cps as the lowest frequency perceived as a tone, if computation were
to be performed using the Organ of Corti as a delay line, then the delay
between the oval window and the termination should be approximately
12 milliseconds (minimum), the duration of a quarter wave of the tone.
If the length is taken as 48 mm, the resultant velocity is only 4 meters
per second, which is slow in most infinite media, but the construction
of the cochlea is such that lower velocities than those in infinite media

are possible (the flexible central membrane as a low modulus element,
for example). Von Bekesy (ref. 2.1) reports that the velocity varies
with position, and that approximately 5 milliseconds delay occurs between
the oval window and a point 34 mm distant. Thus the possibility of
12 milliseconds to termination is reasonable (a linear extrapolation of Fig-
ure 11-53, page 458, implies about 40 mm, ref. 2.1).

With the mechanics of the detector assumed to be suitable to the
mathematical model, it becomes possible to continue the examination from
this viewpoint. We should point out that the theory here presented involves
a time distributed sensor, with anechoic characteristics, feeding a computer
system using only time delays, attenuations and signed additions (plus or
minus). This is different from Ewald's theory as described by Von Bekesy
as "a sinusoidal movement of the stapes sets up a series of standing
waves along the basilar membrane ....". It is essential to our theory
that anechoic properties dominate.

In view of the correlation lengths concerned, 2 to 300 microseconds
for localization, 400 to 2600 microseconds for speech, and 3 to 40 milli-
seconds for reverberation not discernable as separate echoes, it seems
entirely possible that the mathematical functions for attention and recogni-
tion in these domains can be set up in the nervous system directly at the
basilar membrane. Where longer times are concerned, and multiple trunk
or feedback models are examined, the computation arrangement must be
within the ennervating system or in the cortex. It is our present view that
localization, because of its survival function most likely is performed
near the nerve endings.

To construct the function which would provide attention to a particular
location with the Organ of Corti as a delay line, the connections could be
as sketched in Figure 2-5.

$$\sum_{n=0}^{4} \epsilon^{-s(\tau_4 - \tau_n)}$$

$$\tau_4 < 300 \, \mu \, sec$$
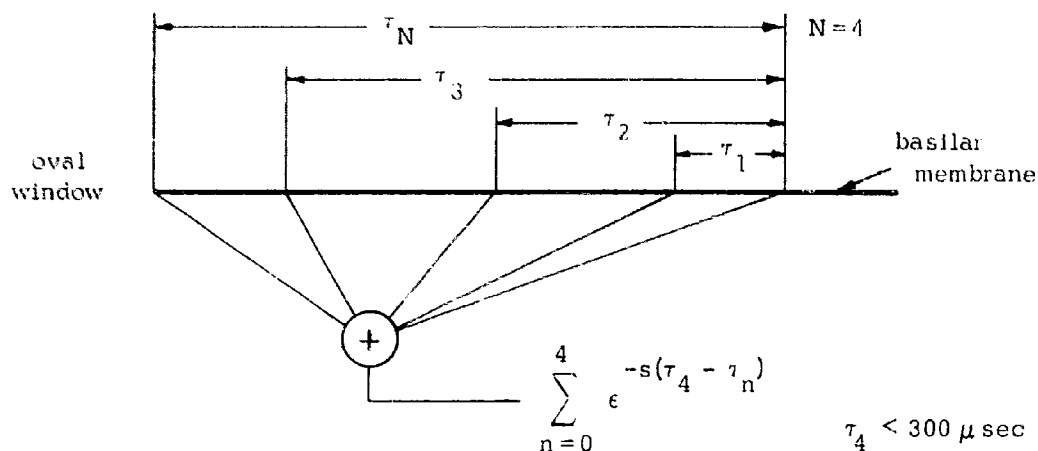
Figure 2-5.   Construction of pinnal attention function at the basilar membrane.

Attention functions may be constructed on any interval along the Organ of Corti which provides the requisite delay lengths, so that other functions may also be provided applicable to reverberation or the character of the sound.

The recognition of a particular pure tone by construction of an attention function suggests an interesting process. If a maximum correlation length is chosen, τ_N, then the correlation sequence can be written as in equation (2.3.1)

$$C_T(s) = \sum_{n=0}^{N} \epsilon^{-s(\tau_N - 2^{-n} \tau_N)}$$

$$= \epsilon^{-s\tau_N} \sum_{n=0}^{N} \epsilon^{+s2^{-n}\tau_N} \qquad (2.3.1)$$

$C_T(s)$ = pure tone correlation function

This equation represents correlation by octaves, and the power $(N-1)$ defines the minimum difference in interval which will correlate, or the pitch of the tone. Consider the definition of a pure tone as the repetition of a single cycle, as given in equation (2.3.2)

$$H(s) = P_T(s) \sum_{n=0}^{\infty} \epsilon^{-sn\tau} \qquad (2.3.2)$$

$$P_T(s) = \text{one cycle pure sine wave of period}$$
$$\tau = 2^{(1-N)} \tau_N$$

Since the function is periodic on the interval $2^{(1-N)} \tau_N$ the result of the correlation applied to the signal is

$$M(s) = (N+1) P_T(s) \qquad (2.3.3)$$

There are several consequences of this consideration:

1. The octave decision process suggests an octave identity in pure tones.

2. Higher frequencies have a greater possible correlation $(N+1)$ for a given total delay length.

3. Resolution of higher frequencies will be poorer than median length frequencies $(\frac{N+1}{N+2}$ approaches 1) for a fixed maximum delay length.

4. Resolution of low frequencies will be poorer than median length frequencies $(N+1$ approaches 1) for a fixed maximum delay length.

One possible consequent adaptive process is the selection of a computational delay length to optimize perception and resolution in the range of consideration. This suggests that continued attention to high frequencies would permit improvement of resolution.

While the octave decision process is one possibility, there are two others. One simply determines the shortest correlation length, or period of the signal; the other proceeds from a minimum length cycle by cycle to a suitable correlation. There are, of course, a large but finite number of ways of performing the measurement to provide both attention and decision. This suggests that the harmonic relations of music may involve processes providing attention and decision in economical ways. The octave attention and decision process is the simplest. Any power of two above the longest interval will correlate in the same arrangement, but the shortest difference will determine the requirement for decision. It is easily observed that ordinarily, in simple harmonies, the highest note provides the melodic line.

2.4   Mechanisms of Perception

When we consider the rationality of the models provided and the measurable behavior consistent with them, we are provoked to inquire into the physiological mechanism, the biophysics as contrasted to mathematics and mechanics. Historically, the electrical potential spikes in the nerves, a consequence of ion concentration shifts, have been considered a likely carrier of the information transmitted by the nerves. However, when we require mathematical functions to be constructed for attention and recognition which require high channel capacity for their performance, and use the distribution of nerve elements in that construct, the capacity of the electrical signals to fulfill the requirement becomes questionable.

We are able to form a hypothesis which will provide consistency by the use of more recent developments in biophysics and physics by assuming that transition of electrons between energy levels in the organic molecules provides information, propagated by the photon emitted.

13

The hypothesis assumes that there are many energy levels, metastable with significant half lifetimes, which can be filled by metabolic processes, and which can be stimulated into transition by any of the sensed phenomena (sound, light, heat, etc.). The transition is made by the emission of a photon, which stimulates transition in adjacent states and thus propagates. The electrical spikes, by hypothesis, are associated with the restoration of the occupance of higher energy levels (as in the laser or maser amplifier) by metabolic process.

While this model remains to be investigated, it provides for high channel capacity, few parallel channels (two are sufficient if the pumping spikes are brief) for continuous operation, encoding of particular source signals, computational networks (for K factors greater than unity), and selectivity to stimulus. It also provides a rationale for the myelin sheath of nerves as a photon path. There are also a number of possibilities of no interest to our present problem, but of general interest; among these are mediation of hormone production by photon catalysis, monocell sensation and computation, spectrum perception by multilevel systems (as in color vision).

2.5 Speech

If we examine the mechanics of speech production, we again find reverberation. The vocal pulse, or aerodynamic noise as in a whisper, provides the stimulus. The organization of the vocal tract performs the transformation. The classical method of speech examination by 'formant frequencies' suffers from two deficiencies.

1. Power density Fourier analysis omits any time dependent relationships.
2. Absolute frequency characterization does not permit scale changes in speech to remain intelligible.

14

The first of the deficiencies is demonstrated by the necessity for simultaneous pulsing of reconstituting filters in vocoder work. The second of the two deficiencies, as shown by speech changes in taped speech, indicates significance or recognition as a dimensionless quantity of relationship and not an absolute dependence on given frequencies. We may resolve these deficiencies by considering the transient character of speech formation, and the construction of dimensionless relationships for recognition.

In the most general representation, equation (2.5.1) applies.

$$H(s) = P(s) \sum_{n=0}^{\infty} a_n \epsilon^{-s\tau_n} \tag{2.5.1}$$

$H(s)$ = speech code element

$P(s)$ = stimulus

a_n = coefficient of reflectivity for the nth delay

τ_n = delay in arrival of the stimulus through the nth path

Although the general expression is necessarily exact, it is not necessarily the most useful. If we consider that the vocal tract of children is smaller than that of the adult and that variations in tape speed over a relatively wide range ($\pm 1.5:1$) retain intelligibility we may inquire concerning the simplest dimensionless characterization. Since the absolute delay is not measurable in the perception in question, and would vary with the distance of the hearer from the talker, we can normalize the absolute representation by considering

$$a_0 = 1 \tag{2.5.2}$$

$$\tau_0 = 0 \tag{2.5.3}$$

The first significant perception is then τ_1 in the normalized signal. In an absolute system, the time between the first arrival and the second, τ_1, could provide a code. In a dimensionless system the ratio between τ_1 and the next longest delay can provide a code. The simplest dimensionless characterization can then be written

$$H(s) = P(s) \sum_{n=0}^{2} a_n \epsilon^{-s\tau_n} \qquad (2.5.4)$$

$$a_0 = 1$$

$$\tau_0 = 0$$

Recognition is provided by ascertaining the relative values of τ_1 and τ_2, as in equation (2.5.5)

$$\rho_{1,2} = \frac{\tau_2}{\tau_1} \qquad (2.5.5)$$

$\rho_{1,2}$ = character of the speech code element

τ_1 = shortest delay

τ_2 = second shortest delay

In equation (2.5.1), H(s) was carefully stated as "speech code element" as contrasted to the linguistic term "phoneme." It is easy to observe, by making a tape loop of a voiced sound such as "ah," that the dynamics of speech are necessary to its ordinary characterization. The tape loop of a voiced sound, when played continually, quickly becomes a machine-like buzz with only traces of the subjective natural sound remaining. We thus assume that 'speech' is produced by sequences of the "speech code elements," consonant sounds being produced by rapid variation and vowel sounds being produced by relatively slow variation. Thus the characterization of the "speech code element" is but the first step in recognition of meaningful speech.

A variety of experiments were undertaken once the hypotheses were formed, the first concerned what could be done to normal speech which would destroy or which would retain the basic coding. If we assume that the vocal pulse is similar to the sketch of Figure 2-6, we can make some predictions.
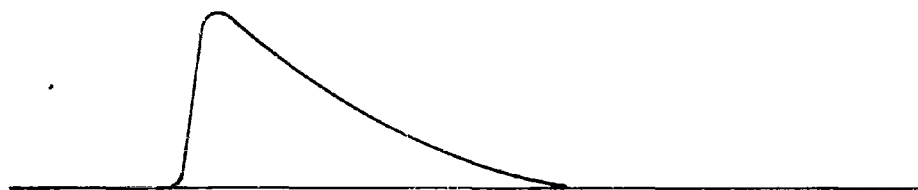
Figure 2-6. An assumed vocal pulse.

This pulse can be approximately described by the following time function, equation (2.5.6)

$$P(t) = U(t) \left[1 - \epsilon^{-k_1 t} \right] \epsilon^{-k_2 t} \qquad (2.5.6)$$

$U(t)$ = unit function

= $0 : t < 0$

= $1 : t > 0$

k_1 = factor of the steep rise

k_2 = factor of the slower fall

If we examine the first derivative the function for $t > 0$, we obtain the following equation, (2.5.7)

$$\frac{dP(t)}{dt} = \epsilon^{-k_2 t} \left[(k_1 + k_2) \epsilon^{-k_1 t} - k_2 \right] \qquad (2.5.7)$$

If we set the derivative equal to zero we obtain equation (2.5.8)

$$(k_1 + k_2) \epsilon^{-k_1 t_{01}} = k_2 \qquad (2.5.8)$$

Now consider the process of taking the derivative of the speech signal, $H(s)$, in equation (2.5.4) to obtain $s\,H(s)$ in equation (2.5.9)

$$s\,H(s) \;=\; s\,P(s) \sum_{n=0}^{2} a_n \epsilon^{-s\tau_n} \tag{2.5.9}$$

It is apparent that taking the derivative following transformation is equivalent to applying a signal equivalent to the derivative of the vocal pulse (necessarily since these are linear transformations). Thus the time derivative of the speech should be intelligible if the characterization is transformation dependent, not structure dependent. Conversely, if the speech is structure dependent, not transform dependent, the result should not be intelligible. (We know this before by whispered speech.) More interesting, the zero of (2.5.8) should also transform the same way. Thus clipping all of the speech but the zero crossings after differentiation should alter the recognition little if the following condition is met, equation (2.5.10)

$$t_{01} \;\ll\; \tau_1 \tag{2.5.10}$$

We also know that the recognition result is obtained (ref. 2.2).

If we examine higher derivatives of the speech signal, we can write a general expression, equation (2.5.11)

$$s^k\,H(s) \;=\; s^k\,P(s) \sum_{n=0}^{2} a_n \epsilon^{-s\tau_n} \tag{2.5.11}$$

In consequence, all orders of derivatives should be intelligible and the increase in zeros of the higher derivatives may contribute to greater redundance in clipped speech recognition. The second derivative was tested and the results are good.

We may now assume that the problem of recognition is to find a recognizing transformation, of the form previously given, which matches

the characterizing transformation of the vocal tract. To do this we need a measure to apply. If we consider the recognition transformation applied to the characterizing transformation, we have equation (2.5.12)

$$A(s) = P(s) \sum_{n=0}^{2} a_n \epsilon^{-s\tau_n} \sum_{k=0}^{2} a_n \epsilon^{-s(\tau_2 - \tau_n)}$$

$$A_R(s) = \epsilon^{-s\tau_2} P(s) \sum_{n=0}^{2} a_n^2 + \sum_{j=0}^{2} \sum_{\substack{k=0 \\ j \neq k}}^{2} a_k a_{0j} \epsilon^{-s\tau_k} \epsilon^{+s\tau_j}$$

$$(2.5.12)$$

$A(s)$ = $H(s)$ followed by a recognition transform

$A_R(s)$ = the form of $A(s)$ upon recognition

To simplify, assume all a_n are unity. The power before recognition is given in equation (2.5.13)

$$U_1 = 3 \times 1^2 = 3 \qquad (2.5.13)$$

U_1 = power before recognition

After application of the recognition function, but before the τ_n match, the power will be given by equation (2.5.14)

$$\overset{\bullet}{U_2} = 9 \times 1^2 = 9 \qquad (2.5.14)$$

U_2 = power after computation but without recognition

When the τ_n match, the power will rise to the value given by equation (2.5.15)

$$U_3 = 3^2 + 6 \times 1^2 = 15 \qquad (2.5.15)$$

To compute the recognition delay values, we need only search for the set of τ_n which changes the resultant signal power ratio for the recognized signal prior to post processing from 3 to 5. There are other methods which have been discussed in prior reports, but this is a simple measuring method, dimensionless on absolute loudness. Even though there is a signal which is louder than the one recognized, the sharp increase in value at recognition should provide knowledge of the basic speech code element. Once the basic element is recognized, the program of symbols to provide meaning may be constructed.

It appears that almost any linear transformation applied to the speech waveform will be recognizable, so long as the time correlation points are recognizable. This may be expressed in equation (2.5.16).

$$T(s)\ H(s)\ =\ T(s)\,P(s)\ \sum_{n=0}^{2} a_n \epsilon^{-s\tau_n} \qquad (2.5.16)$$

Thus this form of recognition is a remarkably reliable one. However, if the equivalent stimulus, $T(s)\,P(s)$, is smoothed, so that sharp location of the delay times is lessened, then recognition will be decreased. Smoothing of waveforms can be produced by phase shift alone. Thus processing of the signal by suitable phase shift only systems to produce smoothing should reduce intelligibility, without alteration of the power density spectrum.

To construct such transformations the following equation may be used, (2.5.17)

$$T_p(s)\ =\ \frac{R(s)}{R(s)\ +\ j\,I(s)}\ -\ \frac{1}{2}$$

$$=\ +\ \frac{1}{2}\left[\frac{R(s)\ -\ j\,I(s)}{R(s)\ +\ j\,I(s)}\right] \qquad (2.5.17)$$

R(s) = real part of characteristic

I(s) = imaginary part of characteristic

$T_p(s)$ = phase shift only transformation

A typical second order circuit for this purpose is shown in Figure 2-7.
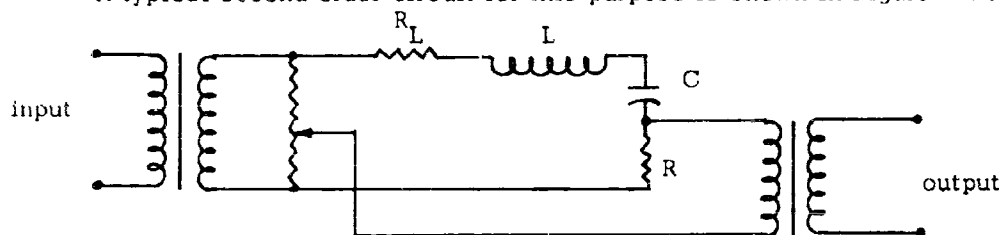
R_L = resistance of inductance L

L = inductor

C = capacitor

R = external resistance

Figure 2-7. A typical second order phase shift only circuit.

When speech is passed through a sequence of such filters to produce phase smoothing, the result can be almost completely unintelligible.

Our experiments also showed that full wave rectified speech is almost completely unintelligible, but that a small imbalance of 3 db in the two sides was sufficient to restore intelligibility somewhat. If we write the transformation in the time domain, as in equation (2.5.18) we can examine the result of full wave rectification.

$$H(t) = \sum_{n=0}^{2} a_n P(t - \tau_n) \qquad (2.5.18)$$

21

Full wave rectification may be written in equation (2.5.19)

$$\left| H(t) \right| = \left| \sum_{n=0}^{2} a_n \, P(t - \tau_n) \right| \qquad (2.5.19)$$

In the electronic amplification of speech signals, conditions are such that equation (2.5.20) must hold.

$$\frac{1}{t_u - t_1} \int_{t_1}^{t_u} H(t) \, dt = 0 \qquad (2.5.20)$$

t_1 = lower bound of sample time

t_u = upper bound of sample time

$t_u - t_1$ = pitch period, or time between vocal pulses.

In this case, equation (2.5.21) is true

$$\left| H(t) \right| \neq \sum_{n=0}^{2} \left| a_n \, P(t - \tau_n) \right| \qquad (2.5.21)$$

Thus the characterizing transformation no longer applies, for

$$\mathcal{L} \sum_{n=0}^{2} \left| a_n \, P(t - \tau_n) \right| = \mathcal{L} \left| P(t) \right| \sum_{n=0}^{2} \left| a_n \right| \epsilon^{-s\tau_n} \qquad (2.5.22)$$

$$\mathcal{L} \triangleq \text{Laplace transformation}$$

and

$$\left| \sum_{n=0}^{2} a_n \, P(t - \tau_n) \right| \neq \sum_{n=0}^{2} \left| a_n \, P(t - \tau_n) \right| \qquad (2.5.23)$$

It is believed, but not demonstrated mathematically that the half wave case preserves the transformation.

In synthesis, we found it possible to produce well recognized voiced sounds by pulse spacing. The initial attempts simply varied the spacing between two pulses as sketched in Figure 2-8.
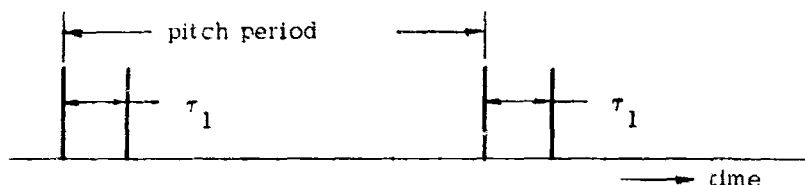
Figure 2-8. Two pulse synthesis of speech sounds.

The two pulse synthesis could be run through the sequence "eeh" to "ah" by varying τ_1. If the spacing were programmed to vary smoothly back and forth, the sound sequence "eeh - ih - eh - ah - eh - ih - eeh" could be repeated, but would not be stationery on any interval. For example, 400 μsec to 1100 μsecs or 800 μsecs to 2200 μsec, or any interval of comparable length in between would sound like the same sequence of speech sounds.

When a dimensioning pulse was placed at 260 μsec, as shown in Figure 2-9, the sequence proceeded from "eeh to ooh" on the interval 400 μsec to 2600 μsec, and the character of the sound would be stationary with respect to the spacing τ_2.

Figure 2-9.   Addition of a dimensioning pulse, τ_1, to the synthesis of speech sound. τ_1 is fixed at 260 μsec, τ_2 is variable 400 μsec to 2600 μsec.

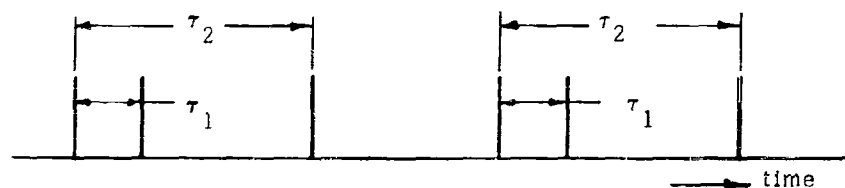Further investigation indicated that addition of redundance in the structure of the pulse, by means of delay line, improved the recognizable character and stationarity of the basic speech code element. The circuitry used is reported elsewhere in this report.

Our experiments generally supported the hypotheses regarding speech formation and recognition, permitting prediction of the effects of processing and also synthesis by single time related systems. The redundance introduced by continued reverberation appears to be simply that, and not essential to the basic code. However, we can write an expression for the reverberant transformation of the vocal tract as equation (2.5.24), using a minimum number of delay lengths (there are undoubtedly more of varying significance).

$$H(s) = \prod_{n=1}^{2} \sum_{k=1}^{\infty} a_n^k \, \epsilon^{-ks\tau_n} \tag{2.5.24}$$

In order to compute, process, or recognize the signal of equation (2.5.24) it is necessary to truncate the expressed infinite series. Practically this amounts to terminating the series when the signal to noise ratio of succeeding terms is so small as to be insignificant. We may then write equation (2.5.25).

$$H(s) = \prod_{n=1}^{2} \sum_{k=0}^{K} a_n^k \, \epsilon^{-sk\tau_n} \tag{2.5.25}$$

A form of recognition function which could be applied to (2.5.25) is given in equation (2.5.26).

$$C(s) = \prod_{n=1}^{2} \sum_{k=0}^{K} a_n^k \, \epsilon^{-s(K\tau_n - k\tau_n)} \tag{2.5.26}$$

Application of C(s) from equation (2.5.26) to the signal H(s) of equation (2.5.25) results in coherent and cross terms, as before, which can be expressed in equation (2.5.27).

$$
H(s)\, C(s) = \prod_{n=1}^{2} \sum_{k=0}^{K} a_n^k\, \epsilon^{-sk\tau_n} \sum_{h=0}^{K} a_n^h\, \epsilon^{-s\,(K\tau_n - h\tau_n)}
$$

$$
= \prod_{n=1}^{2} \epsilon^{-sK\tau_n} \sum_{k=0}^{K} a_n^k\, \epsilon^{-sk\tau_n} \sum_{h=0}^{K} a_n^h\, \epsilon^{+sh\tau_n}
$$

$$
= \epsilon^{-sK\,(\tau_1 + \tau_2)} \left[\sum_{k=0}^{K} \left(a_1^k + a_2^k \right) + \text{cross terms} \right] \tag{2.5.27}
$$

It is assumed that other means of utilizing reverberation in signal improvement and recognition can be found, but the existence of one such function is sufficient to indicate the significance of attention and recognition processes utilizing reverberation.

2.6 Object Recognition

One of the most provocative of the outcomes of the experiment and theory in localization has been the ability to apply the theory consistently to a wide variety of sonic perceptions. It may be observed that localization, speech, room reverberation and music all fall into the class of "recognition," of a place, of a word, of an environment, and of an instrument respectively. Although man does not seem to make full use of the possibilities, bats and dolphins extend the recognition to "kind of object" in food gathering and navigation. Since our work has included interaction with the dolphin, it is appropriate that the present discussion be oriented in that direction.

We assume that the dolphin pulse provides $P(s)$, measured by NOTS to be less than 3 microseconds in rise time, since components in excess of 120 kcps were recorded. If we draw an analogy between the transformation of the vocal pulse by the vocal tract and the transformation of the dolphin pulse by an object, we see that the forms of transformation and recognition functions are identical.

In the case of the dolphin recognition of objects, we can postulate two kinds of transformation, (1) a multiplicity of internal paths and reverberation, and (2) due to a multiplicity of external paths and reverberation. The forms of the equations remain the same in general, but it is useful to introduce the idea of "acoustic coloration," or the effect of materials involved in reflection or transmission on the signal with respect to the Fourier power density spectrum transformation resultant. The two effects are then:

1. Acoustic coloration

2. Time distribution

Both effects may be taken into account in the same mathematical expressions (since they are general), and appear in Laplace transformation notation as in equation (2.6.1)

$$H(s) = P(s) \sum_{n=0}^{\infty} A_n(s) \epsilon^{-s\tau_n} \tag{2.6.1}$$

$A_n(s)$ = coloration transformation through the nth path

τ_n = mean delay through the nth path

The difference between (2.6.1) and previous expressions is the functional character of the coefficient, which previously was written as a constant, indicating coloration in (2.6.1) and no coloration previously. The

equation (2.6.1) is the more general expression more compactly expressed, for A(s) is given by equation (2.6.2)

$$A(s) = \int_0^\infty a(s) \, \epsilon^{-s\tau} \, d\tau \qquad (2.6.2)$$

In equation (2.6.2) the correspondence (2.6.3) may be drawn

$$a(s_1) = a_1 \qquad (2.6.3)$$

Thus a continuous system of delays is implied in (2.6.1) by $A(s)$ accompanied by a discrete system of delays expressed by $\epsilon^{-s\tau_n}$. The continuous system accounts for reflection, absorption, and transmission characteristics of materials. The discrete system accounts for the arrangement of materials into a structure. These expressions thus apply to moths (bats), fish (dolphin), and submarines (man).

The recognition relationships also remain the same, except that orientation of the object in three dimensional space effects the form of the discrete system. If we observe that any two parts of a structure have a maximum separation, when viewed perpendicular to the line joining them, we can write orientation expressions in the discrete system, equation (2.6.4)

$$\tau_n = M^{\tau_n} \cos \theta_n : \quad 0 \leq \theta_n \leq \frac{\pi}{2} \qquad (2.6.4)$$

$$\theta = \quad \text{angle between the connecting line between two structural elements and the normal to the sound wave front.}$$

$$M^{\tau_n} = \quad \text{maximum delay in the nth path between structural elements}$$

It is a trivial observation that in any three dimensional structure not all τ_n can be zero. If we assume a known form transformation, equation (2.6.5)

$$F(s) = \sum_{n=0}^{N} a_n \epsilon^{-s\tau_n} \qquad (2.6.6)$$

Then the rotated form is equation (2.6.7)

$$F_R(s) = \sum_{n=0}^{N} a_n \epsilon^{-s_M\tau_n \cos \theta_n} \qquad (2.6.7)$$

$$F_R(s) = F(s) \text{ transformed by rotations}$$

Given a recognition transformation, equation (2.6.8)

$$C(s) = \sum_{n=0}^{N} a_n \epsilon^{-s(\tau_N - \tau_n)} \qquad (2.6.8)$$

The corresponding rotated form can be given in equation (2.6.9)

$$C_R(s) = \sum_{n=0}^{N} a_n \epsilon^{-s(\tau_N - \tau_n) \cos \theta_n} \qquad (2.6.9)$$

Unfortunately (2.6.9) may have values of anticipation rather than delay, making realization of it not possible in a world bound to the present moment. However, $C_R(s)$ can also be translated in time, as given in equation (2.6.10).

$$_{\tau_c}C_R(s) = \epsilon^{-s\tau_c} \sum_{n=0}^{n} a_n \epsilon^{-s(\tau_N - \tau_n) \cos \theta_n} \qquad (2.6.10)$$

$_{\tau_c}C_R(s) = $ recognition function rotated and translated in time

$\tau_c = $ time translation of recognition

Next it should be observed that only three distinct rotations (or angular degrees of freedom) are possible for a real three dimensional object. The resultant angle of the connecting line on the normal to the wave front must be a consequence of those three rotations, which can be expressed in an object framed coordinate system by equation (2.6.11)

$$\cos \theta_n = R\,(\Psi_n,\, \Phi_1,\, \Phi_2,\, \Phi_3) \qquad (2.6.11)$$

$$\Psi_n \quad = \quad \text{angle of structural line relative to the object frame}$$

$$\Phi_1, \Phi_2, \Phi_3 \quad = \quad \text{rotation of the object frame relative to the observer frame}$$

We can now rewrite equation (2.6.10) as equation (2.6.12)

$$_{\tau_c} C_R(s) = \epsilon^{-s\tau_c} \sum_{n=0}^{N} a_n \epsilon^{-s(\tau_N - \tau_n) R(\Psi_n,\, \Phi_{1,2,3})} \qquad (2.6.12)$$

where τ_N is given by the longest line in the object frame. From equation (2.6.12) it should be possible to recognize an object at any angle and also to determine its structure (internal and external) by viewing it from three different angles (determining τ_n, Ψ_n). All of this can be performed through delays, attenuations, signed additions, and memory. Thus the acoustical recognition by dolphin, given sufficient signal to noise ratios, can provide a detailed, three dimensional model of his surroundings and its contents, or permit him to "see" sonically.

2.7   Localization Experiments

Among the many experiments performed, there are three taken recently in the interest of establishing the accuracy of the system with

the equipment developed, particularly the headphones. The experiment was performed by orienting the pinnae horizontally and producing the sound of a maracas at 16 different positions around the stand. A diagram of the experimental setup is shown in Figure 2-10. In Figure 2-11a, the data is presented on a scatter diagram for elevation. The columns headed "Report" in the figures refer to the reported location by the hearer. The rows labeled "Place" refer to the actual location of the sound source. The consistency of the resultant diagram is indicative of relatively accurate location in elevation. In Figure 2-11b the result of a similar test in azimuth is presented.

In order to separate room effects from the effects of the pinnae, a similar test was made using two bare microphones but otherwise identical equipment. The result is shown in Figure 2-11c.

As is to be expected, the sidedness of the bare microphones is observable, but front-back and up-down ambiguities (equivalent in such a test) show strongly on the scatter diagram. In the case of the microphones with pinnae and the system in use, however, the ambiguities are tremendously reduced, and the position relationships relatively accurately reported.

A last note on these experiments. In the case of localization with pinnae subjects reported confident decisions were easily made. With bare microphones, the expressed subjective confidence in the decision was considerably less.
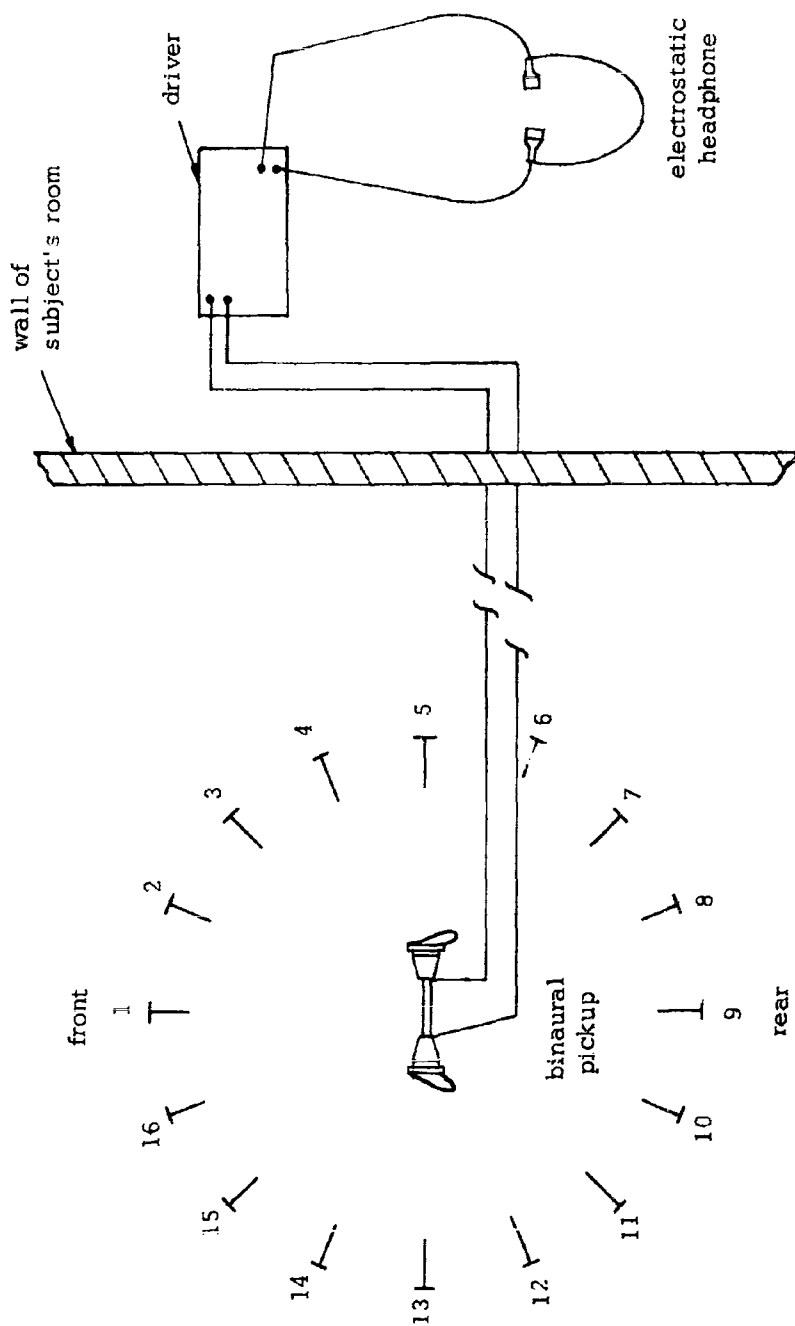
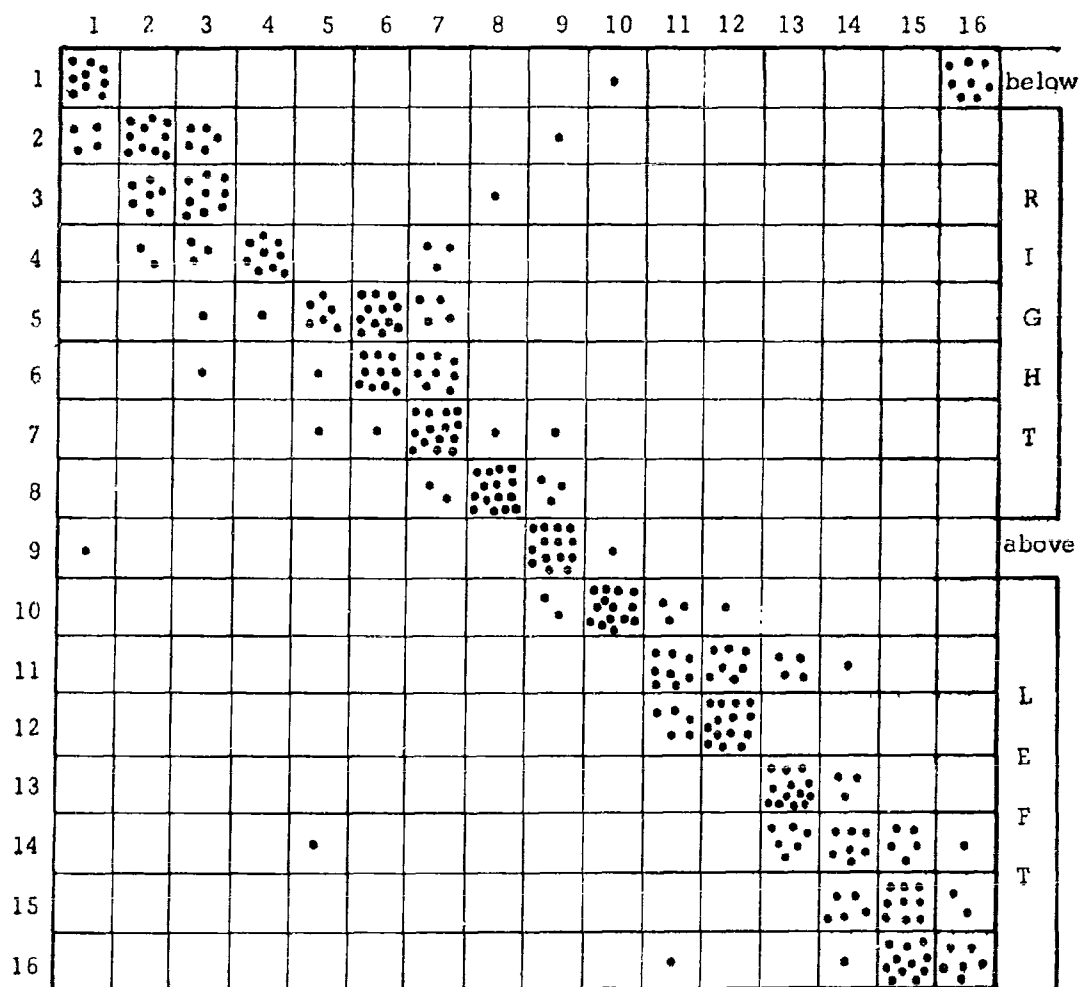Figure 2-10. Setup for localization tests.

REPORT

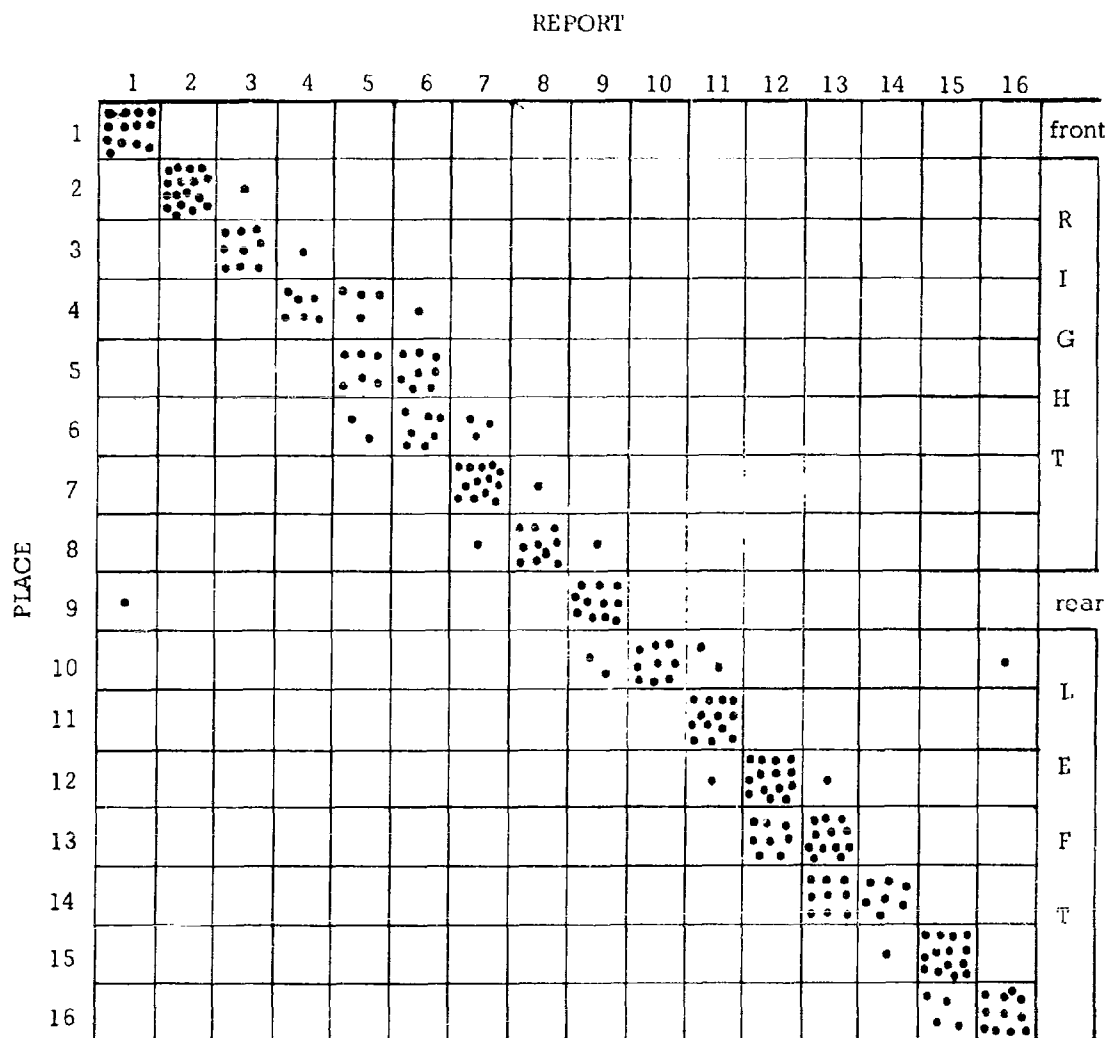Figure 2-11a.   Elevation localization(microphones with pinnae.)

REPORT

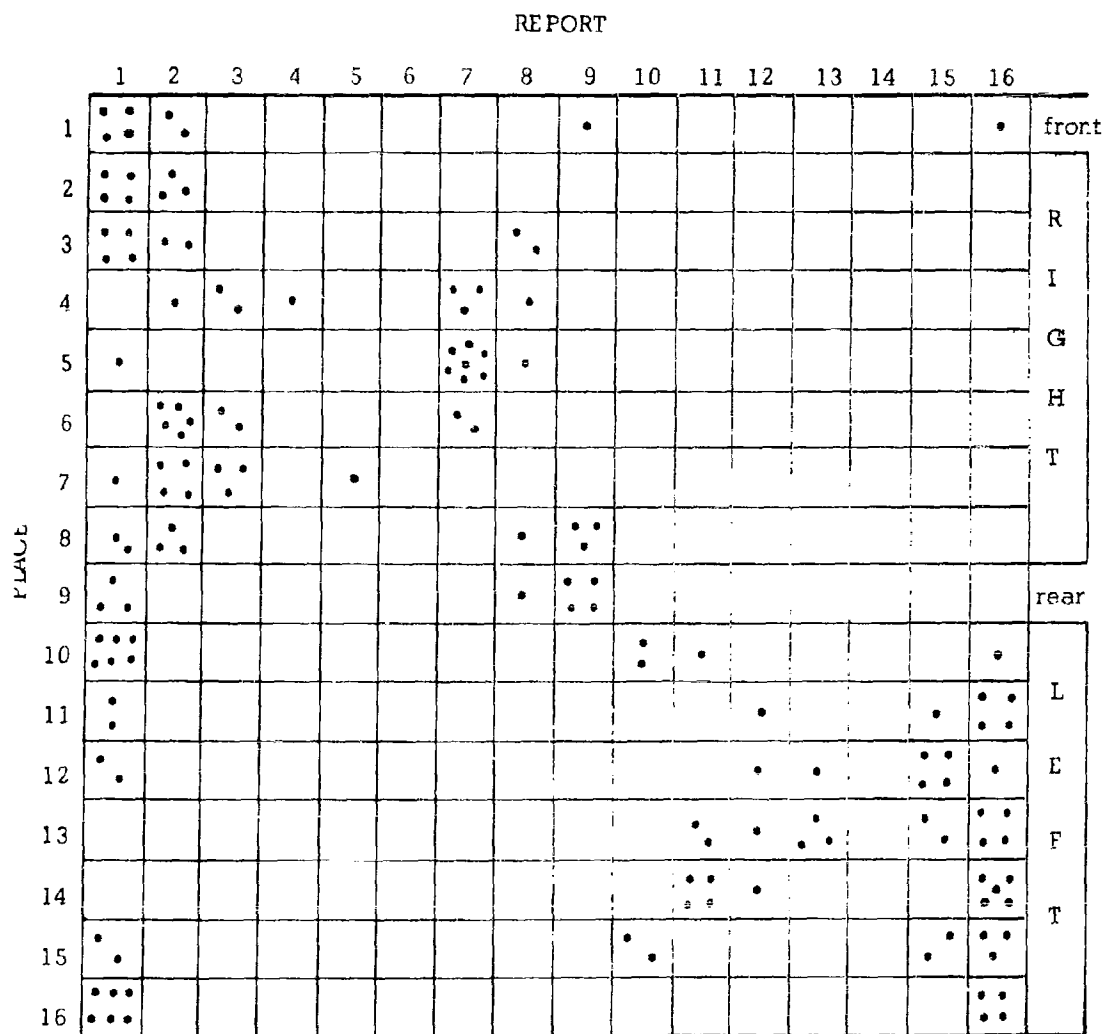Figure 2-11b. Azimuth localization (microphones with pinnae.)

Figure 2-11c.   Localization using bare microphones.

CHAPTER 3

EXPERIMENTAL DEVELOPMENTS

by

Roland L. Plante

3.1   Electrostatic Headphone

The importance of fidelity in all components of a localization system
has been emphasized sufficiently in earlier reports on sound localization
(refs. 3.1, 3.2, 3.3, 3.4). Furthermore, it was stated that the only com-
ponent limiting the total system bandwidth was the headphone. The initial
work to overcome this limitation involved the use of condenser microphones
(Bruel & Kjaer Model 4135) driven as headphones. Localization with this
improvised headset surpassed all expectations. Certain disadvantages
were evident however. The sound level output was only 80 db and the
dynamic range was inadequate. A developmental program was undertaken
therefore to provide electrostatic headphones capable of duplicating the
performance of the improvised B&K headset without the serious drawbacks
noted.

The design goal was to produce a single electrostatic transducer
to reproduce frequencies from 40 cycles per second to 20,000 cycles per
second within a tolerance of $\pm$ 1 db at a sound pressure level of 95 db
re .0002 when loaded into a closed volume approximating the ear canal.
While the design reported here does not meet all the proposed parameters,
it does provide a headset which localization tests have shown to be very
effective in aural coupling (see page 2-28). Furthermore, it embodies
significant improvements over designs reported earlier.

A brief summary of important performance specifications follows:

Diaphragm displacement response: 50-25,000 cycles per second

                                         $\pm$ 3.5 db

Sound pressure in 1 cm^3 coupler   95 db re .0002 μ bar
at 1 kc with signal voltage =
20 v RMS

Bias voltage   200 v DC

The requirement for a constant amplitude-frequency characteristic
was set by consideration of the microphone placement in the pickup. The
microphone is located at the entrance to the ear canal of the replica ear.
Its diaphragm displacement, and hence the electrical output, is propor-
tional to the incident acoustic pressure. To reproduce this pressure
proportionately at the ear canal entrance of the listener requires that the
displacement response of the headphone diaphragm be flat. This does not
imply that the pressure at the ear drum will be constant with frequency.
It is, in fact, anything but flat due to the acoustic properties of the ear
canal and ear drum. Earlier design effort concentrated on producing a
constant acoustic power output, which requires a -20 db/decade diaphragm
displacement function. The headset reported here has a power output
which increases with frequency.

The nominal sound pressure level of 95 db re .0002 μbar is
chosen to provide the dynamic range customarily found in headphones.

3.2   Design

The electrostatic headphone element consists of a fixed back plate
electrode and a metallized mylar diaphragm 0.25 mils thick. Mylar
thickness of 0.15 mils was also used. By designing the element as an
insertion type to be located at the entrance of the ear canal, the diaphragm
displacement requirements to produce adequate sound pressure are lessened.

Furthermore, a flat response characteristic may be expected if the resonant frequency occurs at the upper limit of the desired bandwidth. Then for frequencies lower than the resonant frequency, the displacement, and hence the pressure in a closed cavity (dimensions small compared to wavelength) is constant for constant signal voltage. Since the resonant frequency is heavily damped, it does not appear in the characteristic manner. It was measured by noting the frequency at which $90°$ phase difference between the drive signal and the diaphragm displacement occurs.

Sensitivity of the element depends on the gap between the disphragm and the back plate. The gap is determined by the tension in the diaphragm, the clearance machined into the back plate, and the polarization voltage applied. In the final configuration the operating gap is .00015 inches. This was determined by measuring the capacitance of the element without bias (25 pfd) and with bias (80 pfd). The gap without bias is .0005 inches.

3.3   Performance Measurements

The measurements made to specify the performance of the headphone are: (a) pressure response in a 1 cm^3 coupler; (b) diaphragm displacement response when loaded into a human ear and into the 1 cm^3 coupler; (c) sensitivity; (d) second harmonic distortion; (e) phase response of diaphragm displacement.

Since the headphone element is located at the entrance to the ear canal, it was decided to use as the coupler one whose volume was equal to that of a human ear canal and whose shape would minimize wave length effects. Reference to published data as well as measurement of our own ear canal volume using water and a graduated hypodermic syringe yielded a volume of 1 cm^3. The shape, Figure 3-1, was selected

from reference 3.5 and scaled to give 1 cm^3 cavity volume.

Note: pressure release
hole not shown
.0135" dia. with
8 mil wire inserted

headphone
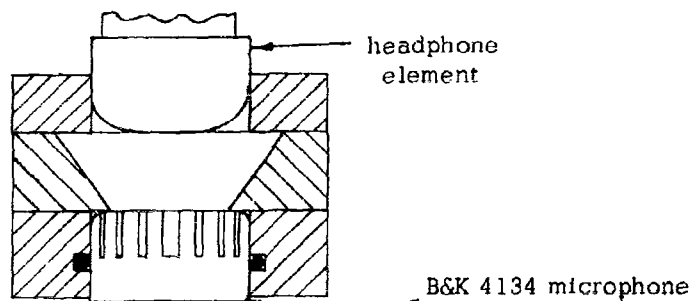element

B&K 4134 microphone

Figure 3-1.   1 cm^3 coupler.

Unfortunately, the exact required shape to give a flat coupler characteristic
was not easily duplicated. One reason was the necessity for retaining
the microphone protector cap. This resulted in a cavity which was shaped
as shown in Figure 3-2 because the diaphragm of the microphone is recessed
from the protector cap.

18.4 mm

9.8mm

6.15 mm

1.2 mm

Note: pressure
release
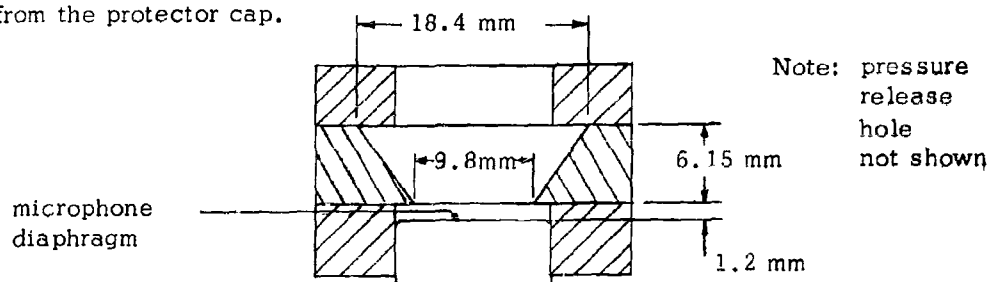hole
not shown

microphone
diaphragm

Figure 3-2. Actual coupler shape caused by retaining
protector cap.

Therefore, the frequency characteristic was measured using
Bruel & Kjaer 4134 microphones as both the transmitter and the receiver.
When this characteristic is subtracted from the pressure response curve
of the headphone in the 1 cm^3 coupler, the displacement response is
obtained. Figure 3-3 shows the coupler characteristics.

The diaphragm displacement was measured using the circuit
shown in Figure 3-4. Operation of the unit is based on varying the capacity

Bruel & Kjaer Microphone Cartridges Model 4134 used as Transmitter and Receiver
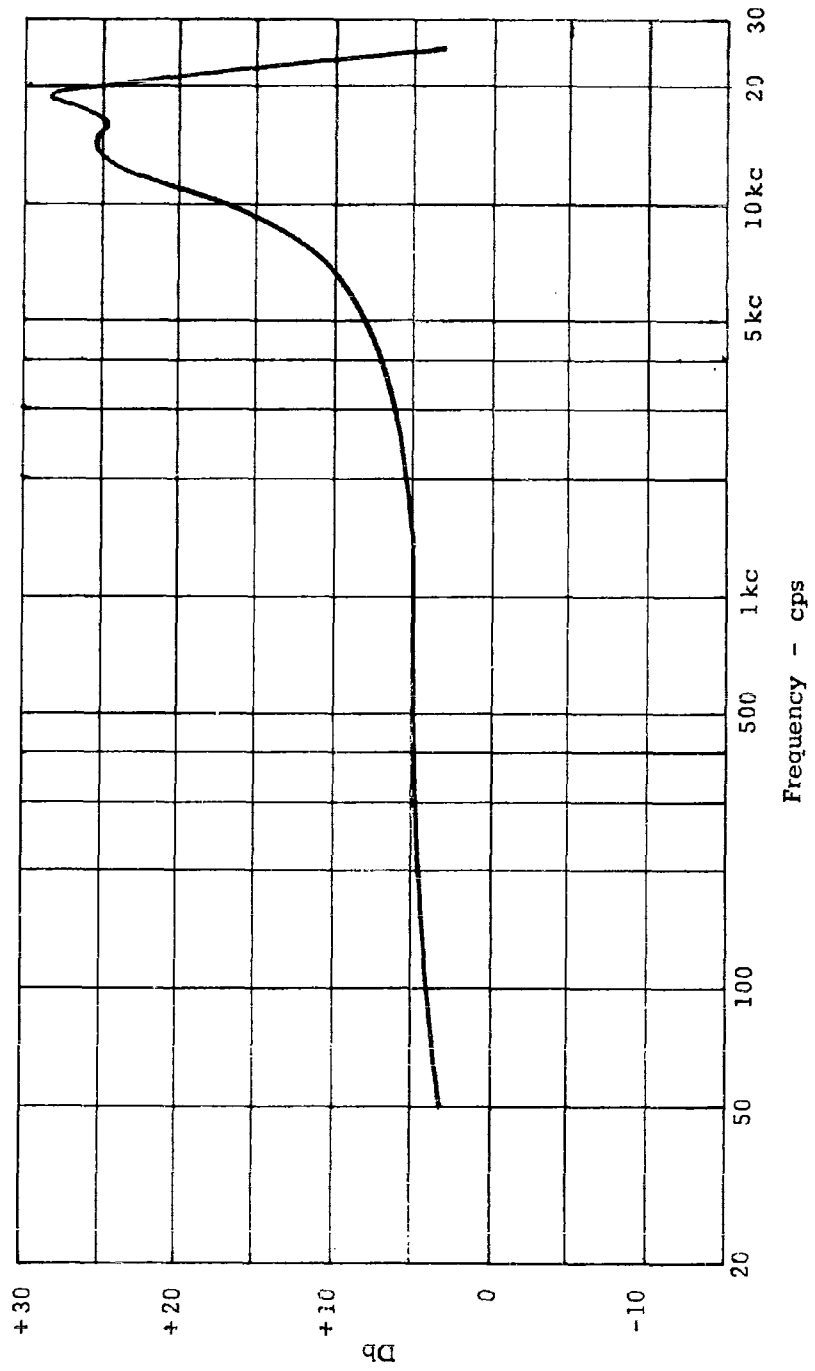
Figure 3-3. 1 cm³ Coupler Characteristic

of a circuit tuned to 21.4 mc. The output is a voltage linearly proportional to the displacement of the diaphragm over the range of interest. Figure 3-5 is the diaphragm displacement response when loaded into the 1 cm^3 coupler and into a human ear canal.

Sensitivity was determined using the 1 cm^3 coupler and a B&K 4134 microphone. By measuring the microphone output and knowing its sensitivity, headphone sensitivity is easily calculated. Figure 3-6 is the pressure response in the coupler. SPL at 1 kc is 95 db re .0002 μ bar.

Second harmonic distortion of the microphone output in the 1 cm^3 coupler was measured using a Hewlett-Packard Wave Analyzer Model 302A. Figure 3-7 shows the measured distortion at 90 db SPL. The distortion shown is due to mechanical and electrical factors. The latter may be calculated.

The electrostatic force on the diaphragm may be written

$$F \sim (E_b + E_s)^2 \qquad (3.3.1)$$

$$E_b = \text{polarization voltage}$$

$$E_s = \text{signal voltage}$$

For a sinusoidal input signal

$$F \sim (E_b + E_p \sin \omega t)^2 \qquad (3.3.2)$$

$$E_p = \text{peak voltage}$$

Expanding equation (3.3.2) and substituting known identities gives equation (3.3.3)

$$F \sim (E_b^2 + \frac{E_p^2}{2}) + 2 E_b E_p \sin \omega t - \frac{E_p^2}{2} \cos 2 \omega t \qquad (3.3.3)$$

The percent second harmonic distortion is

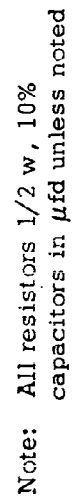$$\% \text{ distortion } = \frac{E_p}{4 E_b} \times 100 \qquad (3.3.4)$$

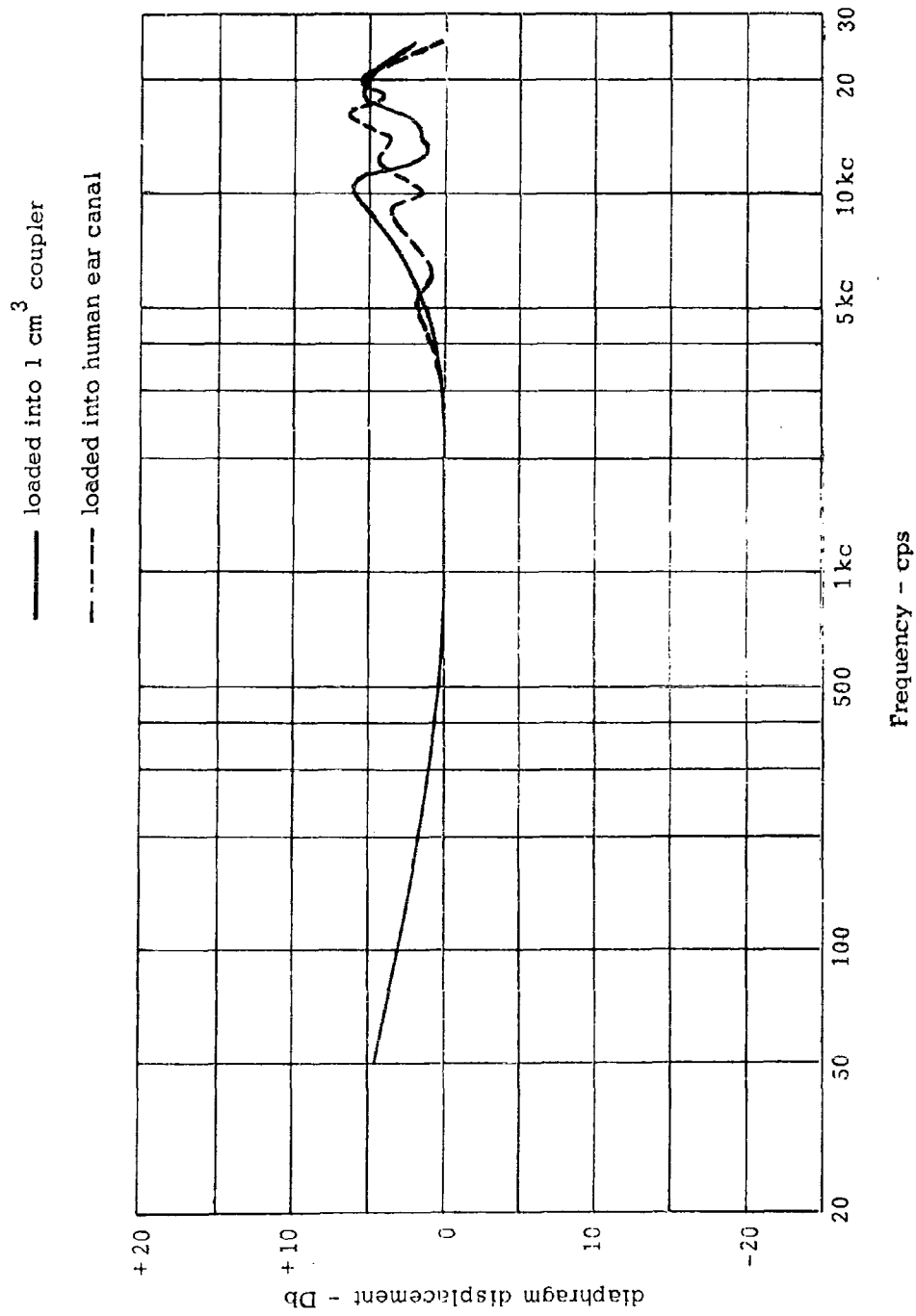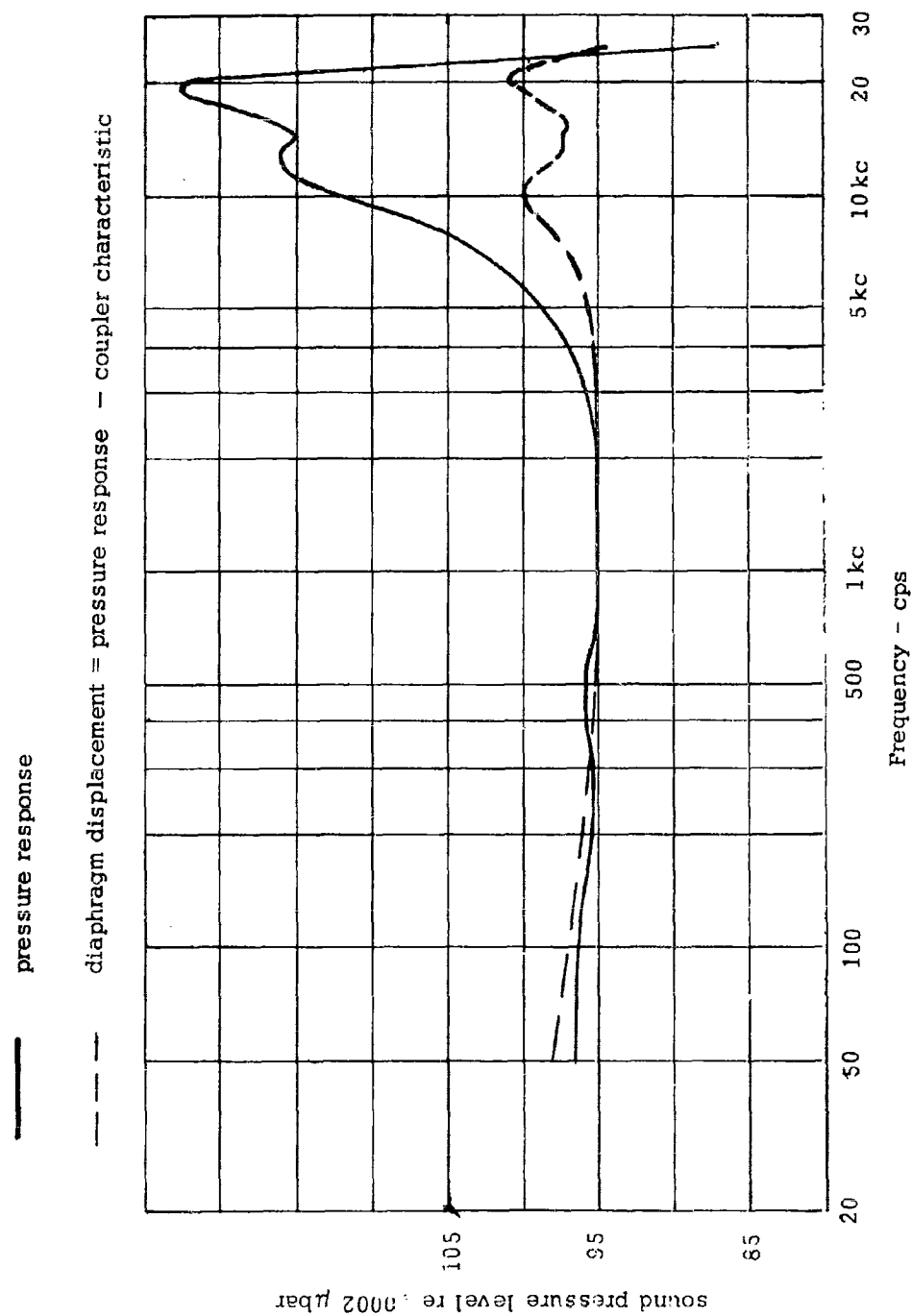Figure 3-4. Circuit used to measure diaphragm displacement.

Note: All resistors 1/2 w, 10%
         capacitors in μfd unless noted

41

Figure 3-5. Diaphragm displacement response.

loaded into 1 cm^3 coupler

loaded into human ear canal

Figure 3-6. Headphone pressure response in 1 cm³ coupler.
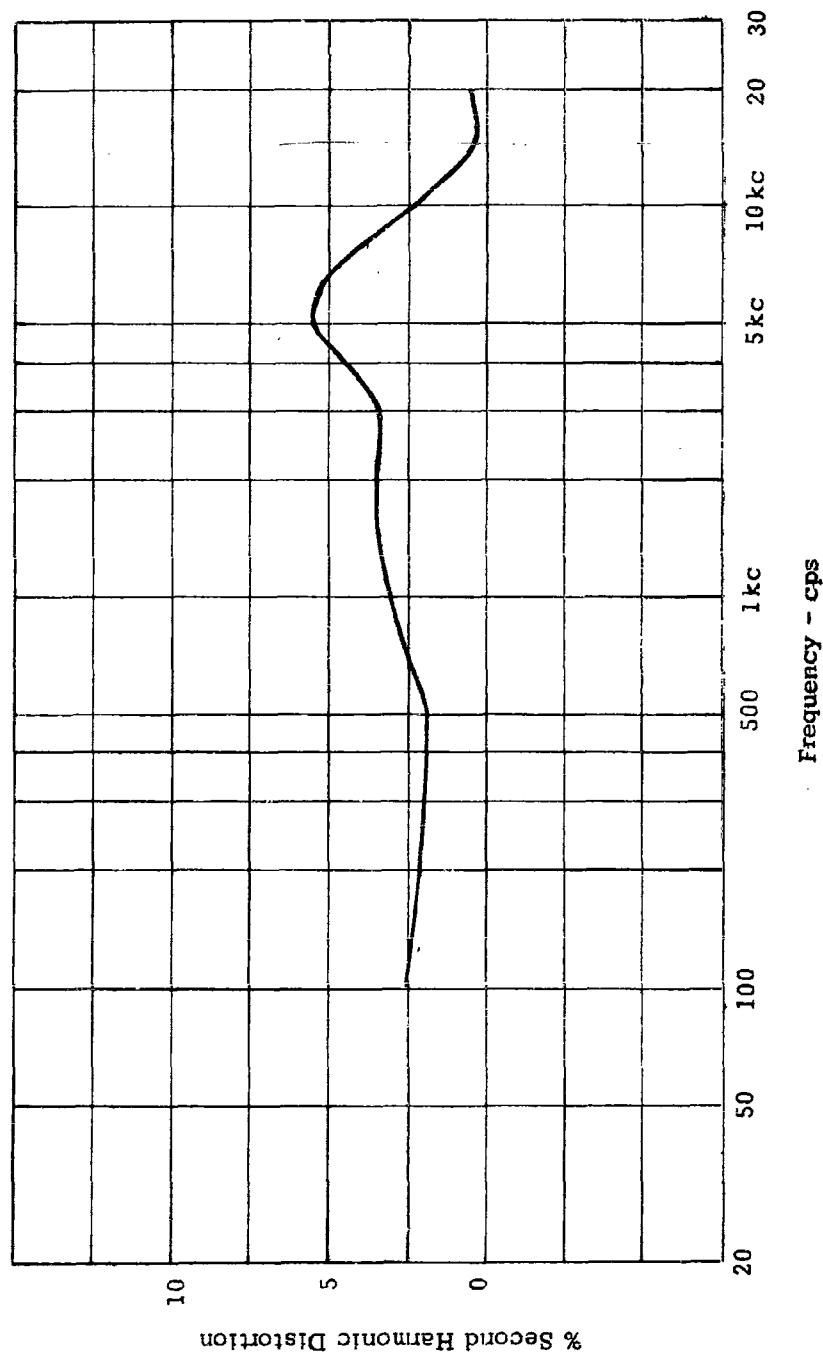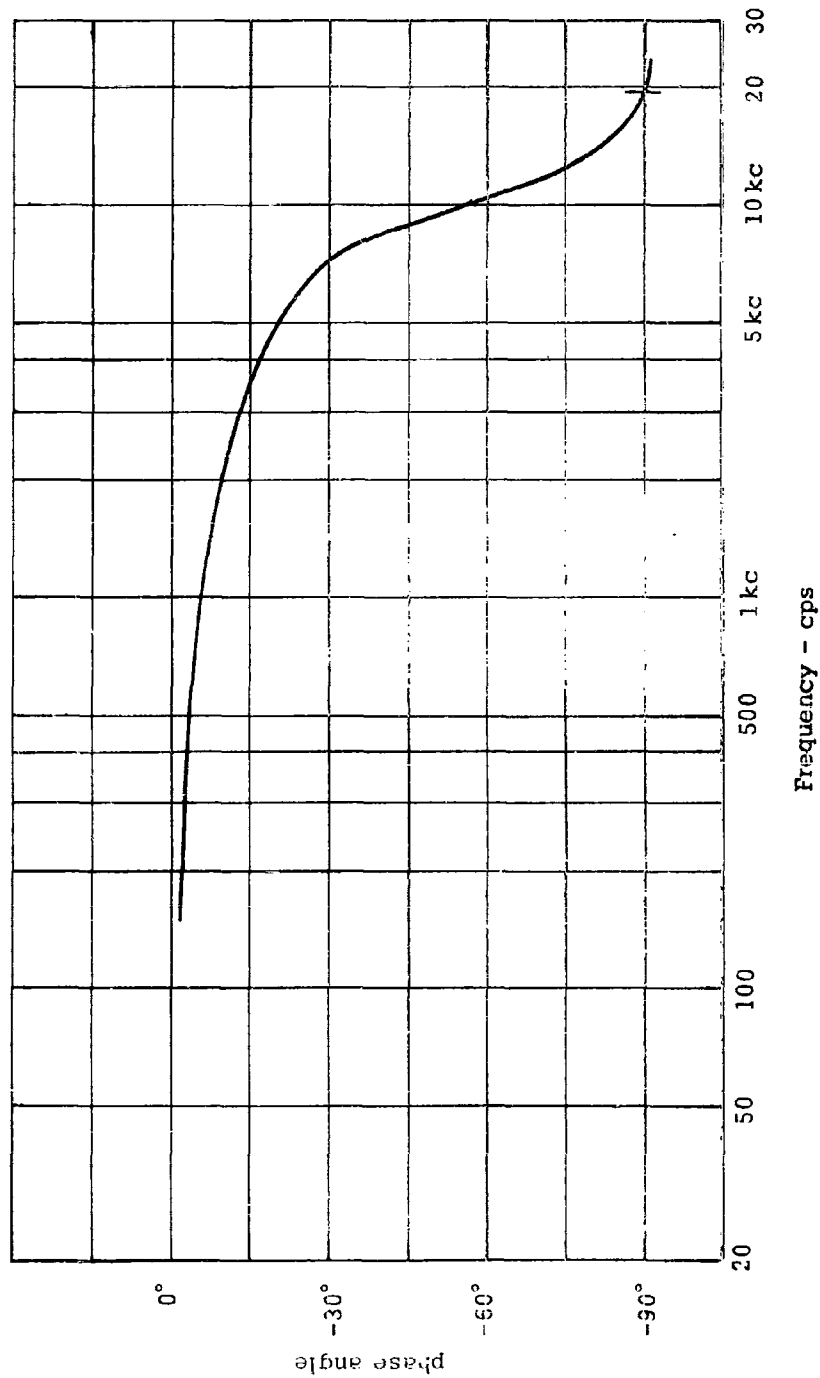
Figure 3-7. Second harmonic distortion.

Figure 3-8. Phase response oh headphone diaphragm.

At 90 db SPL $\quad E_p = 10$ v and $E_b = 200$ v. Therefore, the electrical distortion is 1.25%.

Phase measurements were made with a Tektronix Dual Beam Oscilloscope, Model 502. Figure 3-8 is the measured phase response. Note that 90° phase shift occurs at 19 kc, the resonant frequency.

3.4 Headphone Construction

The configuration of the final element design is shown in Figures 3-9, 3-10, 3-11 and 3-12. With the exception of the relatively high distortion level, (which can be decreased by increase in bias voltage or by increase in sensitivity) the headset shown is eminently suitable to the needs of sound localization.

Proper coupling of the headphone output to the ear canal required the development of suitable ear plugs as shown in Figure 3-12.

While an under-the-chin arrangement is shown, the elements were also adapted to circumaural muffs, in which case each element is spring loaded within the muff.

Some attempts were made to permanently polarize the mylar film so as to eliminate the need for external bias (ref. 3.6). Within the limits of our work, we found that the applied signal voltage tended to reduce the trapped charge, thus reducing the effective bias and increasing the signal distortion. It is recommended that this technique be explored more fully as a means of eliminating what is commonly accepted as the chief drawback to electrostatic headphones, namely, the polarization voltage.
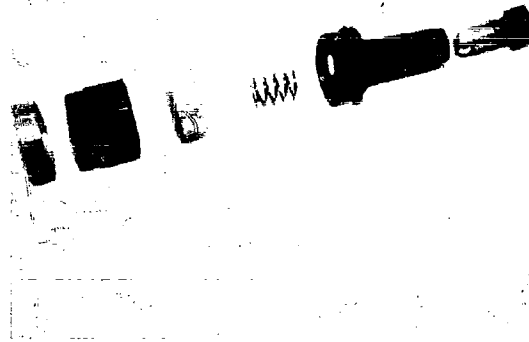
Figure 3-9   Exploded Photo of Element
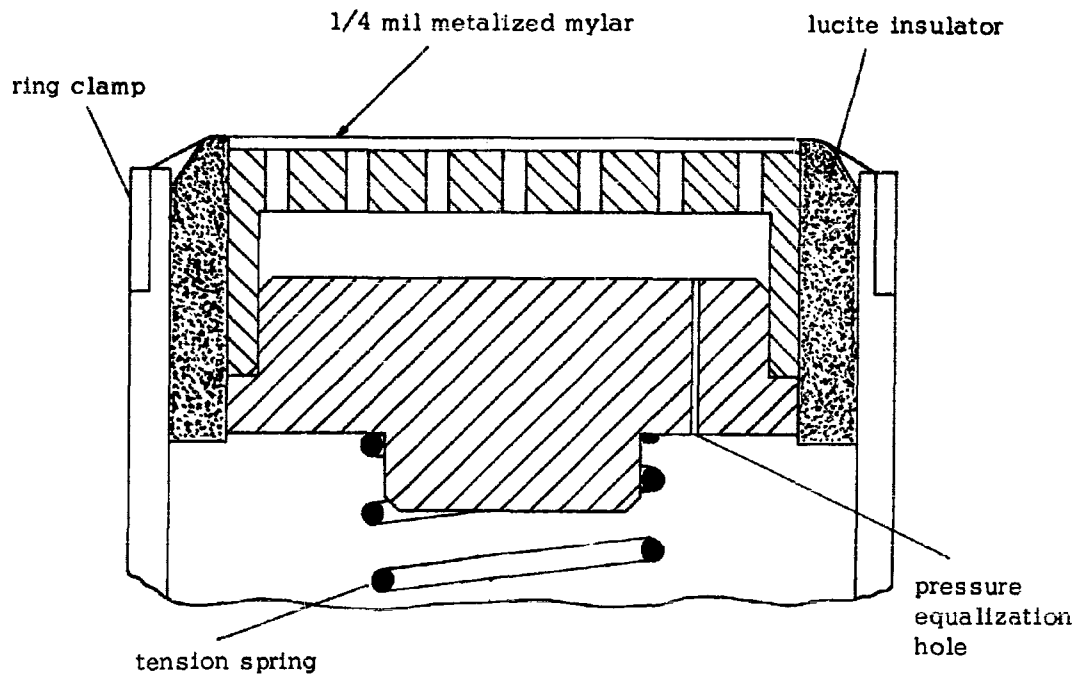

Figure 3-10   Assembled Element

1/4 mil metalized mylar   lucite insulator

ring clamp

pressure
equalization
hole

tension spring

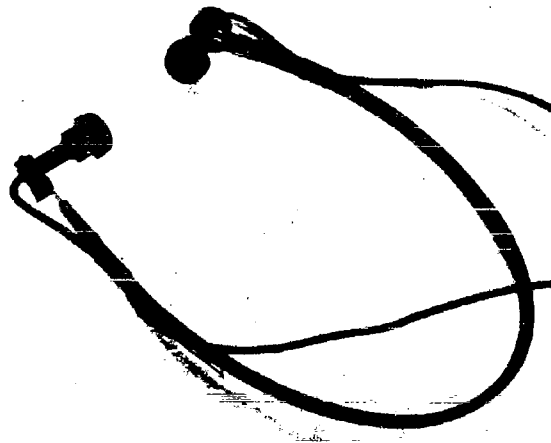Figure 3-11.   Back plate assembly.

Figure 3-12   Electrostatic Headphone

3.5   Headphone Drivers

The amplifier/driver for electrostatic headphones has been dis-
cussed in reference 3.3. Essentially, this amplifier combines a pre-
amplifier function with the drive function supplying the required signal
voltage swing and D.C. polarization.

To make the headphone more compatible with conventional drive
systems, a simple circuit, Figure 3-13, was put together and tested.
This circuit provides compatibility between conventional audio outputs
and the electrostatic headphone.

A third method which was studied and tested is shown in Figure **3-14.**
Here the bias is provided by a capacitor charged to the required voltage
by means of the signal input. This driver eliminates the need for a biasing
battery and provides compatibility with standard audio outputs.

3.6   Acoustic Coloration

Sounds reproduced from the binaural pick-up, Figure 3-15, are
subjectively rich in the mid-frequencies. While this does not appear
to seriously affect the localization transfer capability, it was felt that
a closer reproduction of naturalness would further improve the coupling.

Experiments were designed to study the differences between
human and cast replica ears. A test panel was constructed which
allowed a subject to essentially put his ear on a board, Figure 3-16.
The cast replica could be mounted in the same way, Figure 3-17. The
purpose of the panel was to create equal acoustic conditions surrounding
the ear. A speaker was excited with wide band noise and a probe tube
adapted to a Bruel & Kjaer 4134 microphone. Data was recorded on a
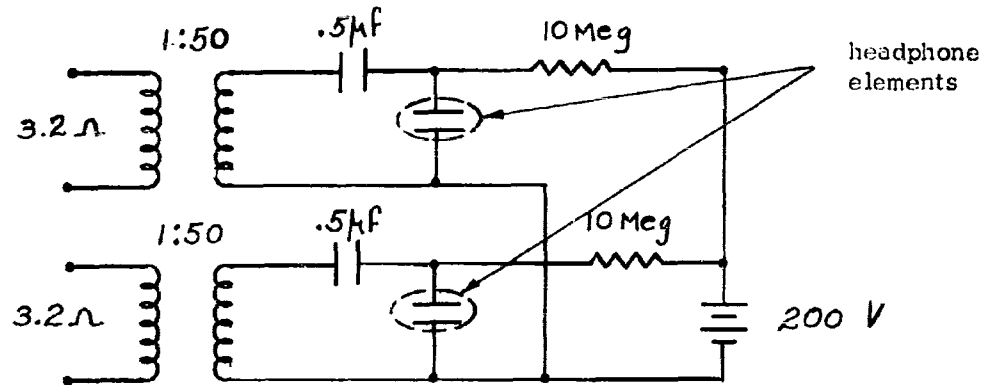B&K Level Recorder fed by a B&K spectrum analyzer.

Figure 3-13.   Electrostatic headphone drive circuit for use
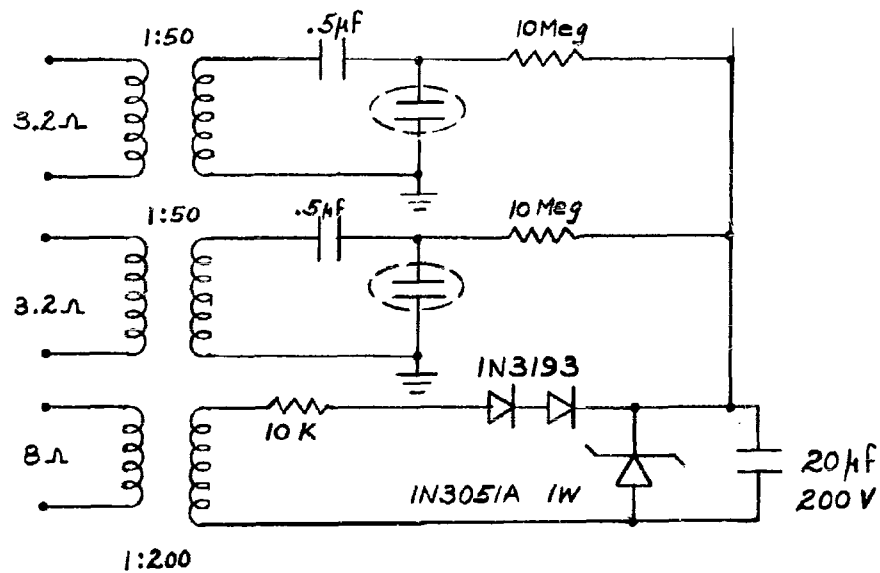with conventional amplifiers.

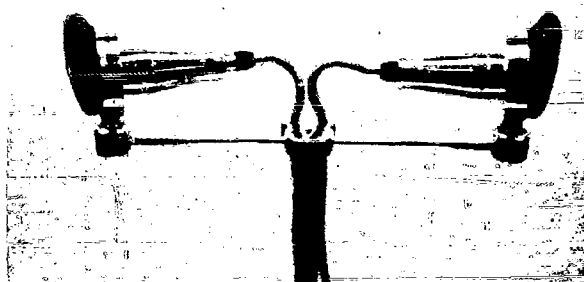Figure 3-14.   Circuit using charged capacitor for bias.

Figure 3-15   Binaural Pickup

Figure 3-16   Human Ear "Mounted"
On Test Panel

Figure 3-17   Ear Replica On Test Panel

Two important measurements were made:

1. The human ear and replica ear were plugged with clay at the canal entrance and probe tube measurements were made on each. No significant difference was noted. It was concluded that the material selected for the ear replica was acoustically satisfactory.

2. The probe tube was placed at the entrance of an open human ear canal and data recorded. The probe was then located at the same position in the replica ear <u>with the microphone in its normal position</u>, i.e., at the entrance to the ear canal. A distinct difference was measured as shown in Figure 3-18. Emphasis of the range between 2 kc and 7 kc is evident. It appears that the energy in this band is naturally coupled to the ear drum. This obviously does not occur in the replica ear-microphone adaptation. Further measurements were made with the microphone in different places without any real improvements.

Two problems remain in the pick-up. One is the proper placement if the microphone which will produce the same characteristics as measured on a human ear and a replica ear. The second is improved acoustic isolation of the microphone in the pick-up for frequencies below 500 cps. Neither of these problems is insurmountable. Continued study should produce a pickup which differs very little with the mechanical function of the external ear.

White noise measured at the entrance of ear canal with
B&K 4134 microphones with probe. Curves shown are
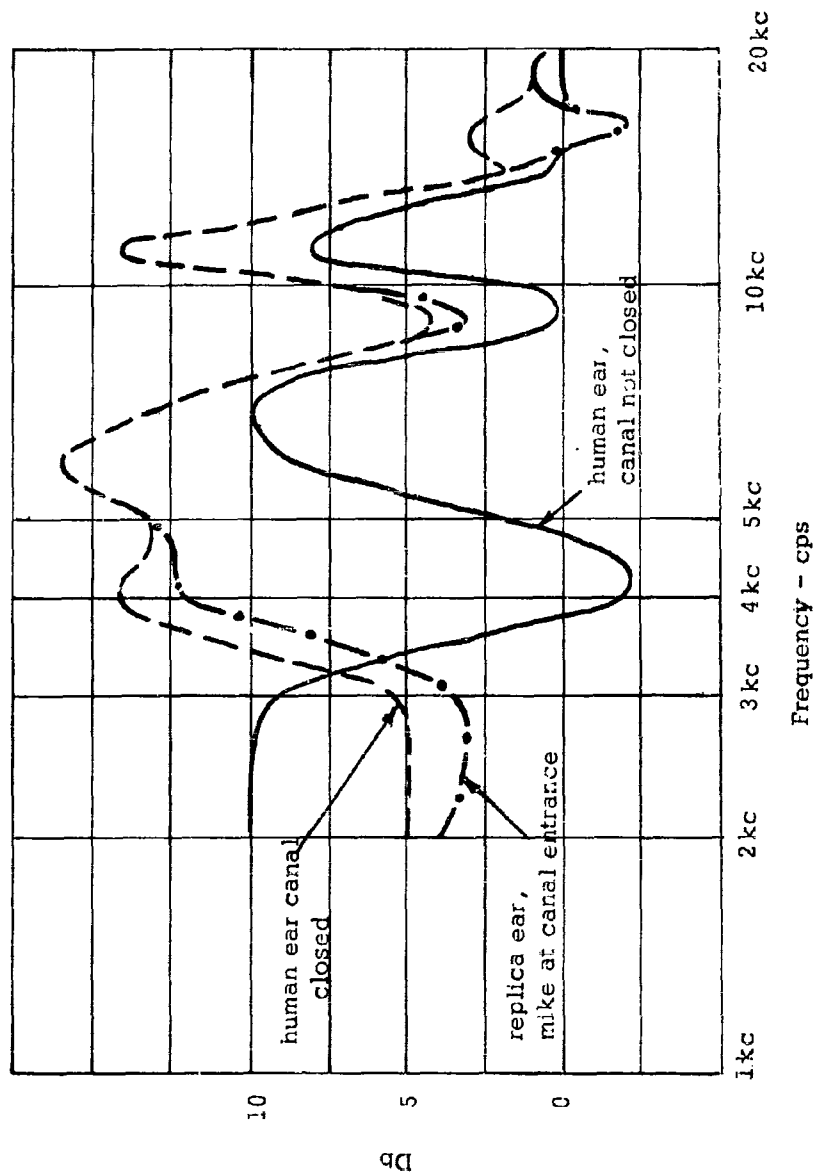the variations around probe tube/noise characteristics.

Figure 3-18. Effects of microphone location in
binaural pickup.

53

CHAPTER 4

INVESTIGATION OF THE RECOGNITION FACTORS

OF HUMAN SPEECH

by

Richard H. Spencer

4.1   Introduction

The theory presented in the preceding chapter was developed from
observations and experiments in auditory localization. Subsequently, it
was used as a basis for investigating the recognition factors of human
speech. The application of the theory to speech work has provided a
new approach to old problems which have yielded little, if any, to the
traditional methods of analysis and experiment.

Interspecies communication between man and porpoise has been
discussed in earlier reports. Under this contract, further improvements
were made in the translators and a meaningful test program was conducted
at Point Mugu, California. Much work remains to be done in this area;
however, it is rewarding that the initial efforts based on the new concepts
have resulted in rapid and significant progress.

4.2   Pitch Extractor

The processing of speech for computer analysis, analysis of the
significance of various parts of the vocal pulse train, digitization of
speech sounds and the processing of speech for speaker identification
clues all frequently require that the start of each voice pulse train be
identified.

Work in the above areas of speech processing and study have in
the past relied on several methods of identifying the start of a vocal
pulse train: (a) Use of a discriminator to identify a predetermined level
of signal; (b) Use of single or double differentiation of the speech
waveform prior to a discriminator; (c) Visual identification from a high
speed recording of the speech waveform. The above methods suffer from
slowness or from uncertainty, particularly if the signal intensity varies.

It has been found that by processing speech waveforms with a
special form of nonlinear circuit a positive identification of the start of
each vocal pulse train is produced. The use of this processing circuit
makes possible quicker, easier and more reliable identification of the
start of the vocal pulse train than is had with the prior methods.

Figure 4-1 illustrates a block diagram of the system by which the
identification of the start of the vocal pulse train is made.

A speech waveform is fed to two peak detectors — one a positive
peak detector, the other a negative peak detector. These two detectors
have carefully selected decay times. The outputs of the two detectors
are weighted and summed. The resultant waveform is a single pulse,
the leading edge of which identifies the start of the vocal pulse train as
indicated in Figure 4-2.

Performance of the pitch extractor is improved by the addition of
a trigger circuit activated by the output waveform. In addition, pre-filtering
of the incoming speech waveform by a low-pass filter provides more
certitude in the output response.

4.3   Gating of Speech Waveforms

In work directed toward determining the significant characteristics
of speech, an experiment was set up which permitted the selective gating
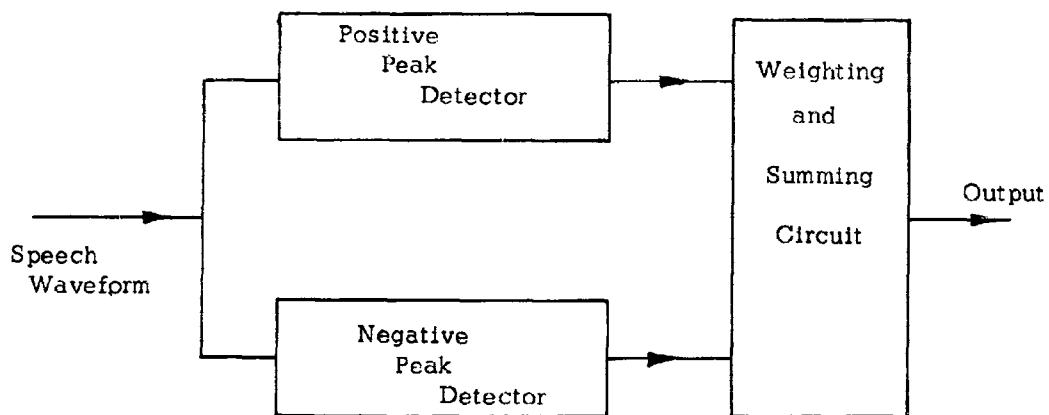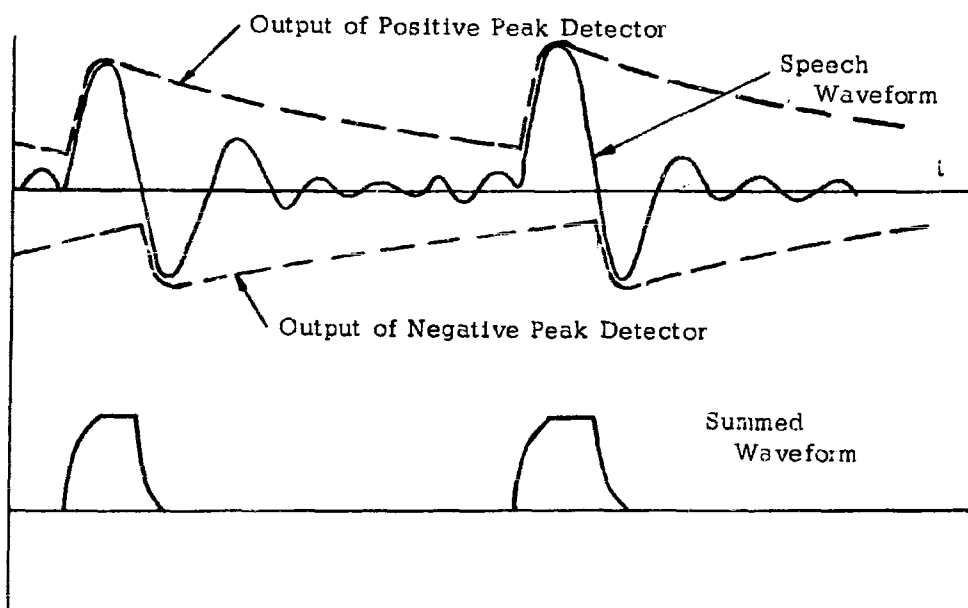
Figure 4-1   Block Diagram of Pitch Extractor

Figure 4-2   Waveforms of Pitch Extractor

of various parts of the vocal pulse train. In this work it was possible
to select a given interval of time at a predetermined time from the vocal
pulse for gating through of the vocal pulse train.

Several tentative conclusions are made on the basis of this work.
(1) With a normal 10 millisecond pitch period any 3 millisecond interval
of "on" time produced intelligible output.

(2) Operation with single or double differentiation of the incoming
speech waveform produced better performance than was shown for unprocessed
speech.

(3) If the gate was programmed to operate on alternate vocal pulse
trains, the subjective interpretation was that the speaker had slowed
down his speech.

4.4   Delay Line Synthesizer

The 100 section, 1000 microsecond delay previously constructed
for NOTS was modified by the addition of 4 compensated level restoring
amplifiers and 5 summing amplifiers of controllable gain and delay posi-
tioning. These modifications provided an instrument for use in analysis
and synthesis of complex waveforms.

4.5   Rasp Generator

The Rasp Generator is a device which converts human speech into
a special form believed useful in communicating with porpoises. Input
to the device is supplied to an AKG capacitor microphone and preamp.
The resulting electrical waveform is amplified, double differentiated in
a delay line circuit, amplified again, then fed to a Schmitt trigger. The
resulting waveform is available directly, integrated once, and integrated
twice. The instrument is portable and operated from rechargeable batteries.
A circuit diagram for this instrument is shown in Figure 4-3.
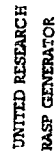
Fig. 4-3 Circuit Diagram of Rasp Generator

4.6   Headset Amplifier

A headset of prior design and construction was modified to improve the gain and phase response. The amplifier is a two-channel, portable, battery-operated unit. The set is intended to be fed from hydrophones with output to a pair of headphones. The enclosure is water resistant, batteries are provided with a built-in recharging circuit. The following specifications are met.

Number of channels:	2
Voltage gain:	variable to 700
Bandwidth: (see Figure 4-4)	20 cps to 70 kcps
Phase shift:	(see Figure 4-5)
$+10°$ to $-10°$	100 to 15,000 cps
$+20°$ to $-20°$	50 to 25,000 cps
$+40°$ to $-40°$	30 to 50,000 cps
Input resistance:	4000 ohms (midband)
Output resistance:	300 ohms
Output swing:	
open circuit:	3 volts rms
470 ohm load:	2.5 volts rms
Noise referred to input:	
input open:	4 microvolts rms
input shorted:	1 microvolt rms
Crosstalk at 1000 cps:	
left to right:	-47 db
right to left:	-43 db
Batteries:	2 CD28 nickel cadmium rechargeable units, .225 ampere hour
Recharge:	22.5 milliamperes per battery by transformer and rectifier operated from line voltage (14 hours for full recharge)

Drain:	30 milliamperes per battery
Operating time between charging:	7 hours
Input connectors:	XLR 3-13 connector
Output connectors:	microphone dual jacks — fit small plugs only
Recommended headphones:	AKG – 400 ohms
Circuit diagram:	See Figure 4-6

4.7   Investigation of Torsional Delay Lines

Two commercial models of sonic torsional delay lines were investigated as a means to achieve multiple delay and gain synthesis of complex signals.

Pulsed sine wave excitation of the lines was used. Output was observed on a CRO. A single delay magnet was installed to produce one delayed signal. The signal produced by this magnet was clearly discernible and could be positioned at any delay within the range of the line. (Induced voltage mode of line operation.)

Two serious drawbacks are evident for the proposed application.

(1) The signal level produced by the delay magnet is no more than 10 times the background noise level even when narrow band amplification is used.

(2) Undesired residual signals occur. These signals are apparently caused by remanent fields left in the torsional line by previous application of a delay magnet. These signals are as much as 1/3 the amplitude of the signal produced by a delay magnet.

Work was suspended on this type of synthesizer because of these two drawbacks.
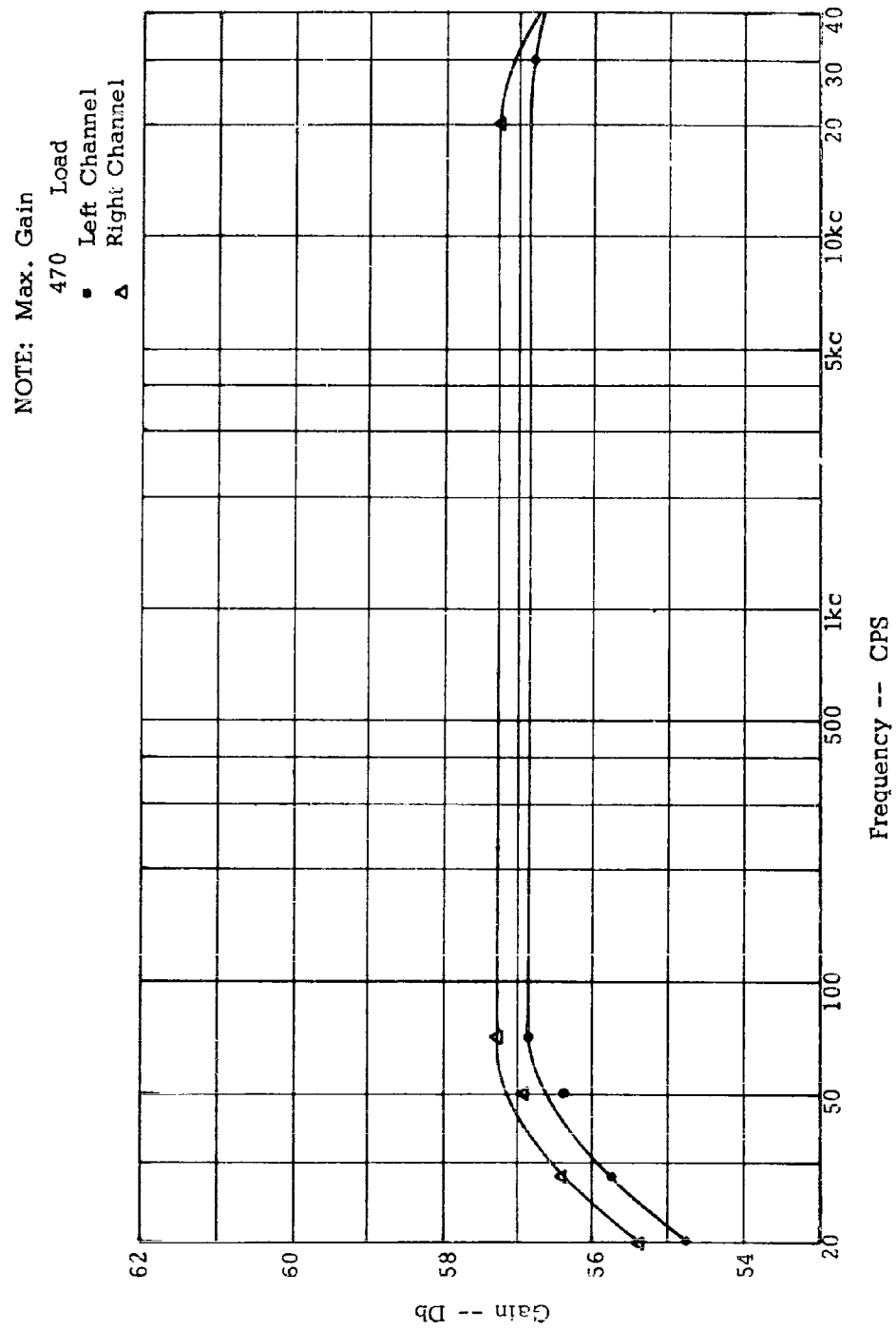
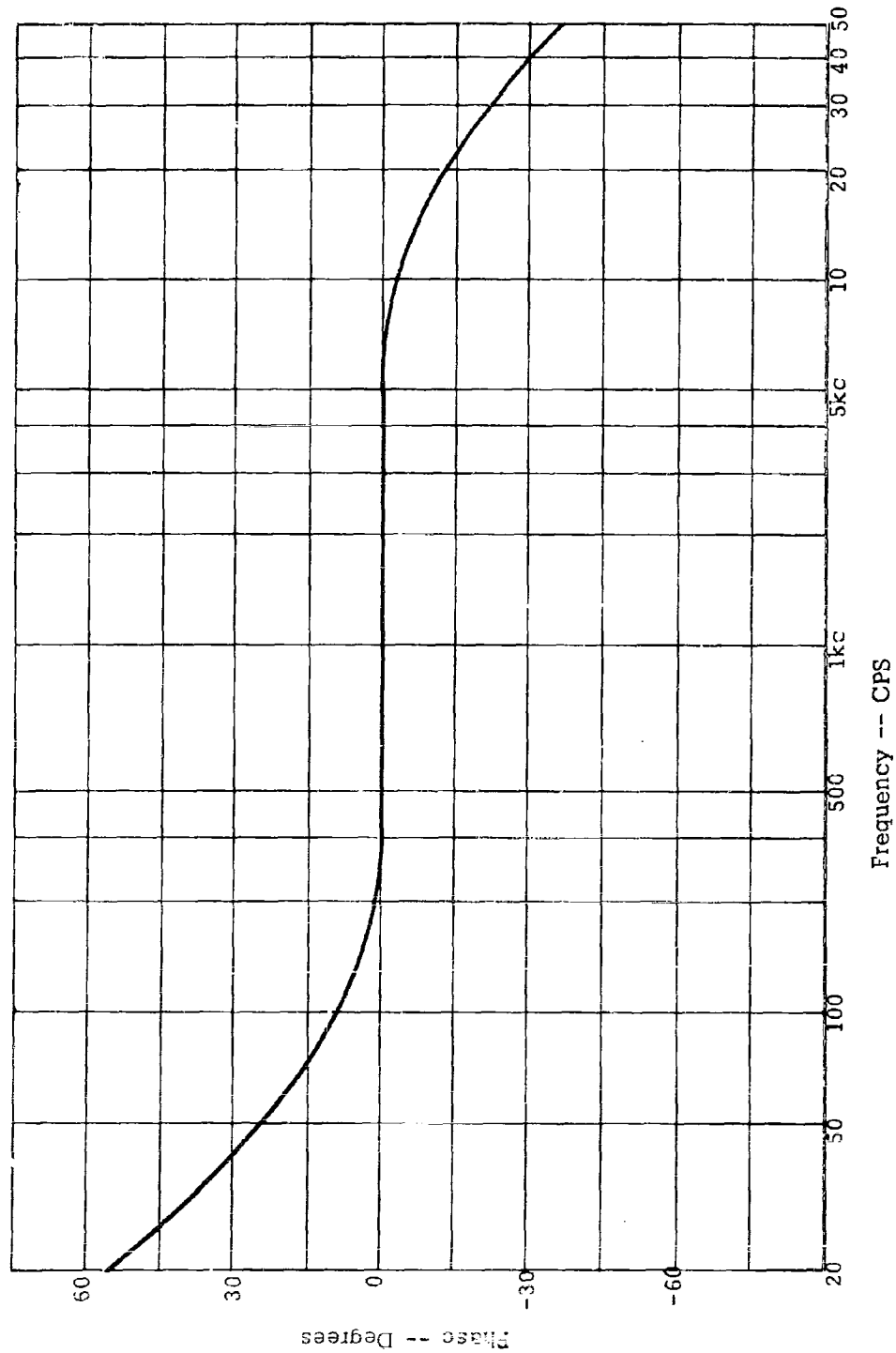Figure 4-4   Headset Amplifier -- Frequency Response

Frequency -- CPS

Phase -- Degrees

Figure 4-5   Headset Amplifier -- Phase Response

2N2374   2N591   2N591   2N591   2-2N1374

LEFT

DUPLICATE OF ABOVE
FOR RIGHT

+12v

25v
FIL. TRANS.   1N3193   0,1   .39

10   .39
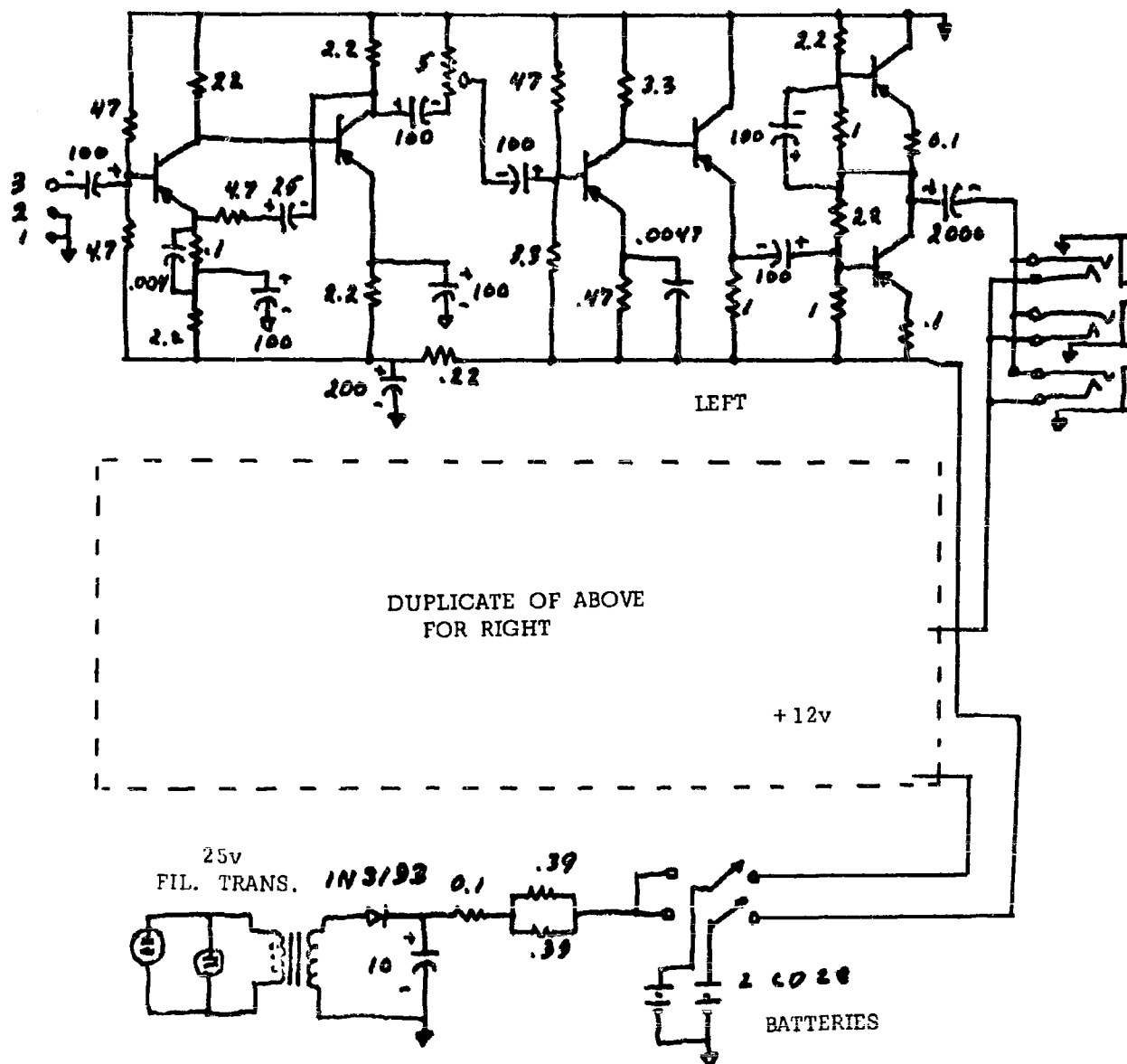
2 C028

BATTERIES

UNITED RESEARCH

HEADSET AMPLIFIER

RESISTANCE IN KILO-OHMS

CAPACITANCE IN MICROFARADS

Fig. 4-6   Circuit Diagram of Headset Amplifier

4.8 Experimental Voice Synthesizer

An early model voice synthesizer which was originally constructed in haywire form was reconstructed into a neat package for use in experiments on phoneme synthesis. The circuit diagram for this device is shown in Figure 4-7.

4.9 Experiments with Voice Boxes

A number of experiments were conducted in attempts to determine the significant variations that occur in the human vocal apparatus in the production of speech sounds. Some of these experiments were concerned with the production of synthetic speech sounds by means of artificial vocal cavities. These artificial vocal cavities were constructed of wood and each had a single movable control piece. Tests were conducted by feeding a short acoustic pulse 100 times per second into the box and listening to the sound emitted from an opening in the box. The sound resulting in these experiments did show phoneme-like character. As in previous experiments on the generation of synthetic speech sounds it was found that dynamics play an important part. If the control piece is moved continuously between its limits it is much easier to identify each sound with a particular phoneme than it is with the control piece in a fixed position. Further, the sounds have a monotonous character; variation of the repetition rate and the exciting pulse would probably remove this undesirable character.

4.10 Photoelectric Waveform Synthesis

A series of experiments was conducted in an investigation of means of producing phoneme-like sounds. Three separate photoelectric techniques were explored in this work.

(1) Use of a template placed in front of a CRO screen together with a viewing photocell and feedback to the CRO. This scheme causes
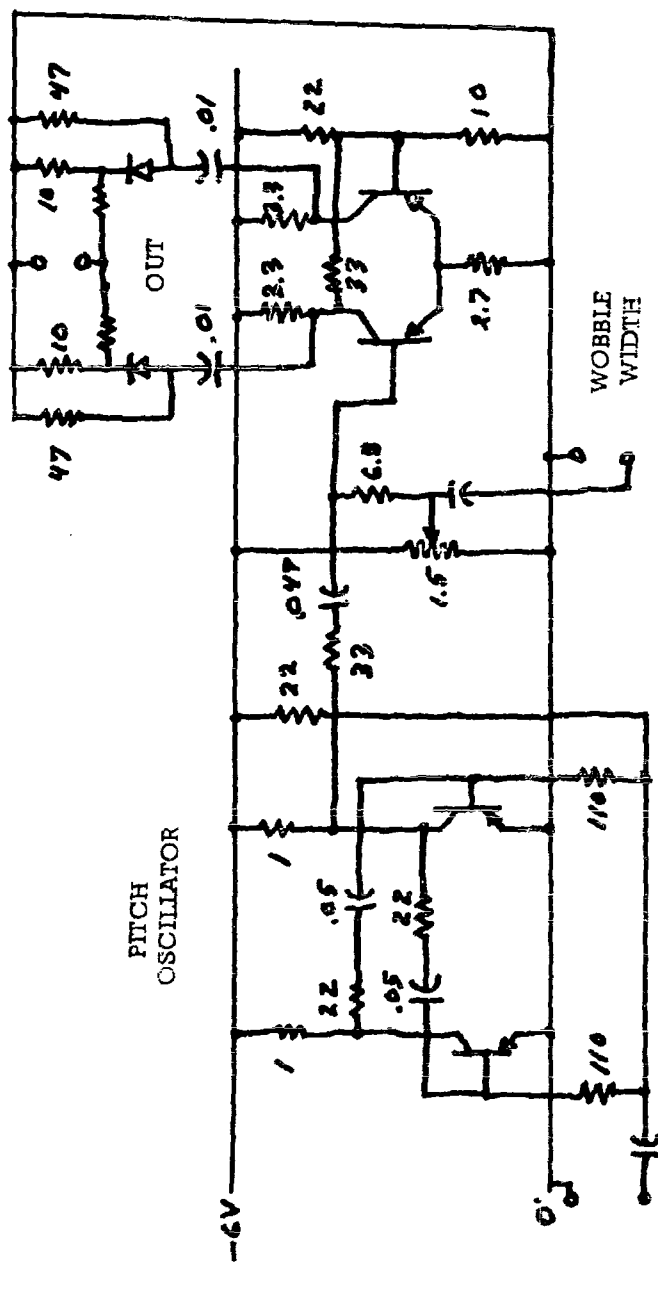
UNITED RESEARCH

PHONEME SYNTHESIZER

RESISTANCE IN KILO-OHMS

CAPACITANCE IN MICROFARADS

PNP TRANSISTORS 2N404

DIODES 1N3193

Fig. 4-7   Circuit Diagram of Phoneme Synthesizer

the CRT beam to position itself at the edge of the template. If the CRT beam is swept by the horizontal deflection circuit of the CRO, the beam follows the edge of the template. The signal produced by the photocell is a replica in time of the waveform represented by the template. The template used was made up of 20 rods each individually adjustable.

(2) Use of an opaque mask on a CRO screen together with a viewing photocell. In this arrangement, with the CRO beam swept across the screen by the deflection circuitry of the CRO, the light reaching the photocell would be blanked out as the beam passed under opaque regions of the mask. The photocell output thus is a signal of one of two values.

(3) Use of a graded mask. A graded mask was made in much the same way that audio signals are recorded in motion picture films. The amplitude-time history of a typical phoneme waveform during one pitch period was recorded from an actual speaker. This amplitude time history was translated into film density versus distance along the film by a photographic process. The graded mask so made was placed in front of a CRO screen. When the CRT beam was caused to sweep across the CRT, the light transmitted by the mask to a photocell caused the photocell output to reproduce the amplitude-time history of the original speaker.

Several conclusions were reached from this experimental work. The subjective impression of sounds obtained by feeding headphones or speakers with the waveforms generated as described above was substantially the same as for sounds generated by the methods discussed in other sections of this report. The sounds are monotonous, discrimination among phoneme sounds is greatly improved by dynamics, that is, changing from sound to sound rather than letting one sound be continuously repeated. No single method produced a significantly better sound than any of the others. As is pointed out in the theory section some randomness and variability in the repetition rate and the fine structure of phoneme waveform is required to provide the quality of realism.

4.11   Double Differentiator

A double differentiator circuit for speech signals was designed and constructed. The circuit is similar to that used in the Rasp Generator. A 12 db per octave shape is shown from 300 cps to 30 kc. The circuit is shown in Figure 4-8.

4.12   Three-Element Synthesizer

An electronic speech sound synthesizer was designed and constructed. The circuit diagram of Figure 4-9 illustrates the functions of the synthesizer. The repetition rate oscillator is an unsymmetric square wave generator of voltage controllable frequency. The output of this oscillator starts each of the interval oscillators in synchronism with the repetition rate oscillator and keeps these oscillators operating during the positive part of the square wave. These interval oscillators are also voltage controllable so that the interval varies in accordance with the control signal.

The repetition rate oscillator also feeds the three exponential envelope generators. Each generator produces a decaying exponential wave of manually adjustable time constant. These exponential envelopes are generated in synchronism with the output of the repetition rate oscillator.

The output of each exponential generator is mixed with the output of an interval oscillator. The resulting waveform is an exponentially decaying square wave pulse.

Outputs of the three signal mixers are summed with adjustable gain into a common output.

The net output is a complex wave made up of three exponentially decaying square wave trains repeatedly generated at a voltage programmable

UNITED RESEARCH
DOUBLE DIFFERENTIATOR

RESISTANCE IN KILO—OHMS
CAPACITANCE IN MICROFARADS
TRANSISTORS 2N404

Fig. 4-8   Circuit Diagram of Double Differentiator

repetition frequency. The intervals of the interval oscillates are each voltage programmable and the decay time constants may be manually adjusted.

4.13   Phase-Lock Whistle Trackers

One of the problems met in the porpoise translator work has been the presence in the water of sounds other than the desired porpoise whistles. Porpoise rasps and clicks and other spurious sounds frequently cause undesired response from the porpoise-man translators.

It appeared promising to design, construct and test whistle tracking circuits which would eliminate much of the undesired background noise. To this end two tracking circuits were designed, fabricated and tested. Both of these circuits operate as phase locked tracking loops. In a phase lock loop, the frequency of an oscillator is controlled such that its frequency is the same as that of an input frequency. This synchronism is obtained by a phase detector fed by the oscillator and by the incoming signal. The output of this detector is a signal which is proportional to the phase difference between the two signals. This phase "error" signal is filtered, then used to control the frequency of the variable oscillator. Once synchronism has been obtained, the variable oscillator automatically tracks the input oscillator in frequency.

The difficulty associated with using a phase lock tracker on porpoise whistles is that of obtaining the initial synchronism. This problem is overcome in the circuits constructed by providing a continuous sweep of the voltage controlled oscillator (VCO) such that it ranges from 4 kc to 16 kc 20 times per second. Synchronism occurs when the VCO frequency is the same as the incoming (porpoise whistle) frequency and the sweep action is automatically overridden. This mode of operation means that there is a delay time between onset of a porpoise whistle and
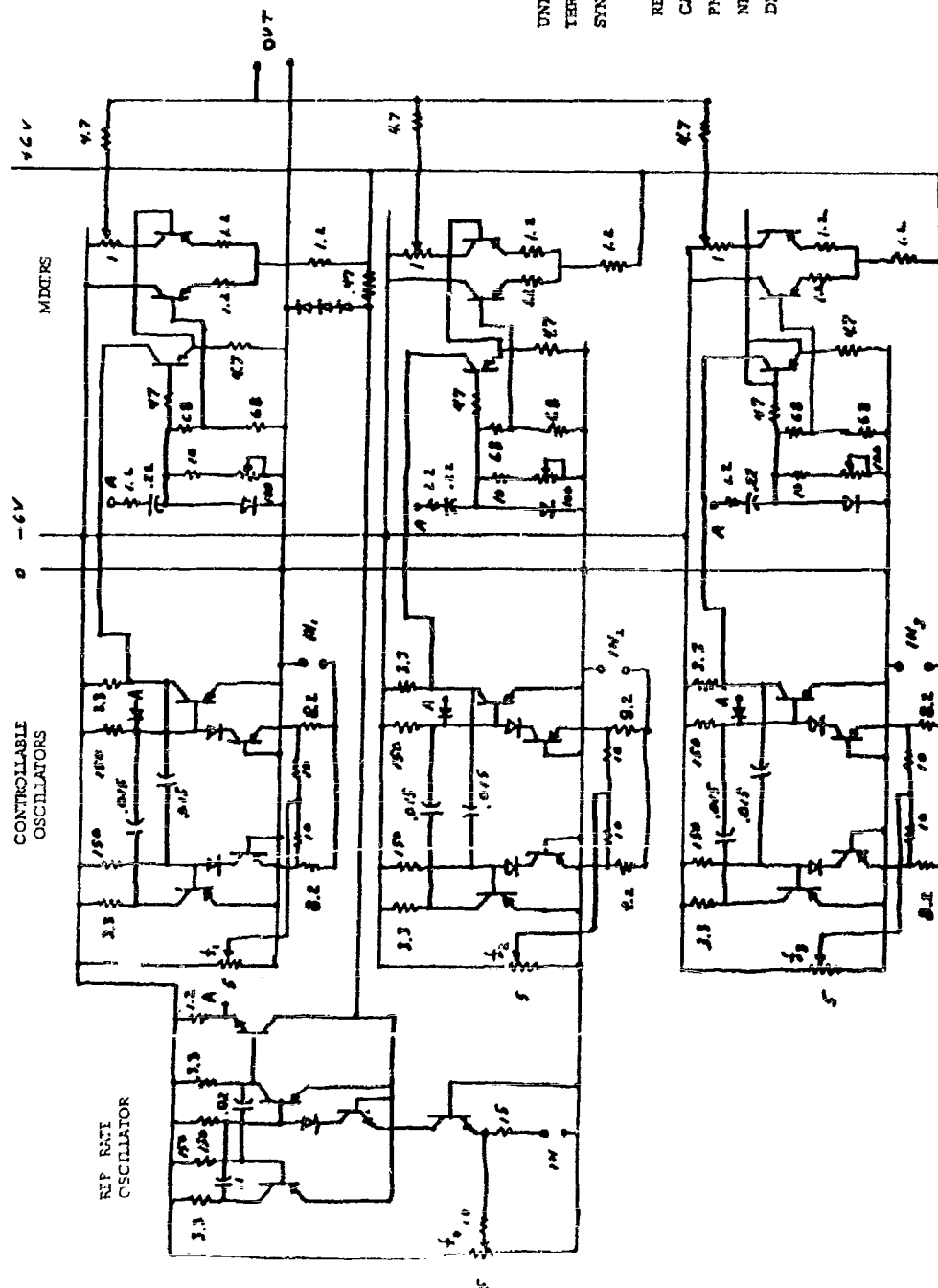
Fig. 4-9   Circuit Diagram of Three Element Synthesizer

the time that the phase lock circuit locks onto the porpoise whistle.
A further delay occurs because it is desirable to gate off the output signal
until lock-on occurs. This function is obtained by using the sweep fre-
quency signal occurring at the phase detector output to operate a gate.

Figures 4-10 and 4-11 are circuit diagrams of the two trackers;
one produces a sine wave output, the other a square wave. In the square
wave circuit the VCO is a square wave oscillator, while in the sine wave
circuit a beat frequency oscillator is used.

Tests showed better performance for the square wave system.
This better performance is attributed to the better dynamic behavior of
the VCO used. That is, the bandwidth, for control purposes, of the
square wave oscillator is wider and hence lends itself better to use in
closed loop operation.

4.14   Mod III Translators

Based on the earlier work on man-to-porpoise and porpoise-to-
man translators, it was decided to design and construct a new translator
system. This new translator would overcome some of the deficiencies of
the prior system and some of the awkwardness of the prior system operation.
The objectives sought in this work were:

(1) A self contained system. All components, preamps, power
amplifiers, batteries, etc. would be in one package.

(2) The instrument would be rechargeable battery operated.

(3) Improvements would be made in the translators based on the
prior experience.

Unfortunately, it was necessary to suspend work on this project
prior to completion. However, the essential circuits have been designed,

constructed and debugged. Remaining to be done is the design and
construction of the battery charging circuits and the packaging of the
entire system,

Figures 4-12 to 4-16 are circuit diagrams of the translators.

4.15 Tests were conducted at Pt. Mugu in late August using the equip-
ment described in reference 3.4. A meta-language was developed, the
purpose of which is to permit man vocalizations to be translated to what
is considered meaningful modulated whistles. The word list to be used
in interaction studies was finalized, as follows:

1. bieib	5. beaeb	9. baiab
2. biaib	6. beaib	10. baieb
3. bieab	7. beieb	11. baeab
4. biaeb	8. beiab	12. baeib

Tapes made show well defined rapid sonic response to input vocalizations.
Initially, two words, "beieb" and "beiab" were used and imitation of each
obtained from the dolphin. The usage by the dolphin was not systematic.
Subsequently, interaction was conducted in three words, "beieb," "beiab"
and "baiab". Imitation was obtained for the first two but not for "baiab".

A review of the tapes made at Pt. Mugu led to the consideration
of the following letters for present and future use in the meta-language:
b, e, a, i, w, y, r. The verification world used in the tests should now be
spelled "biyib" and the negation word spelled "bayal". Tentative construc-
tion for new words, subject to laboratory tests, include the following:

1. rayeb	5. waeb	9. yareb
2. rib	6. wib	10. yib
3. raib	7. wayeb	11. yarib
4. arib	8. awib	12. ayib

Fig. 4-10   Circuit Diagram of Square Wave Porpoise Whistle Tracker
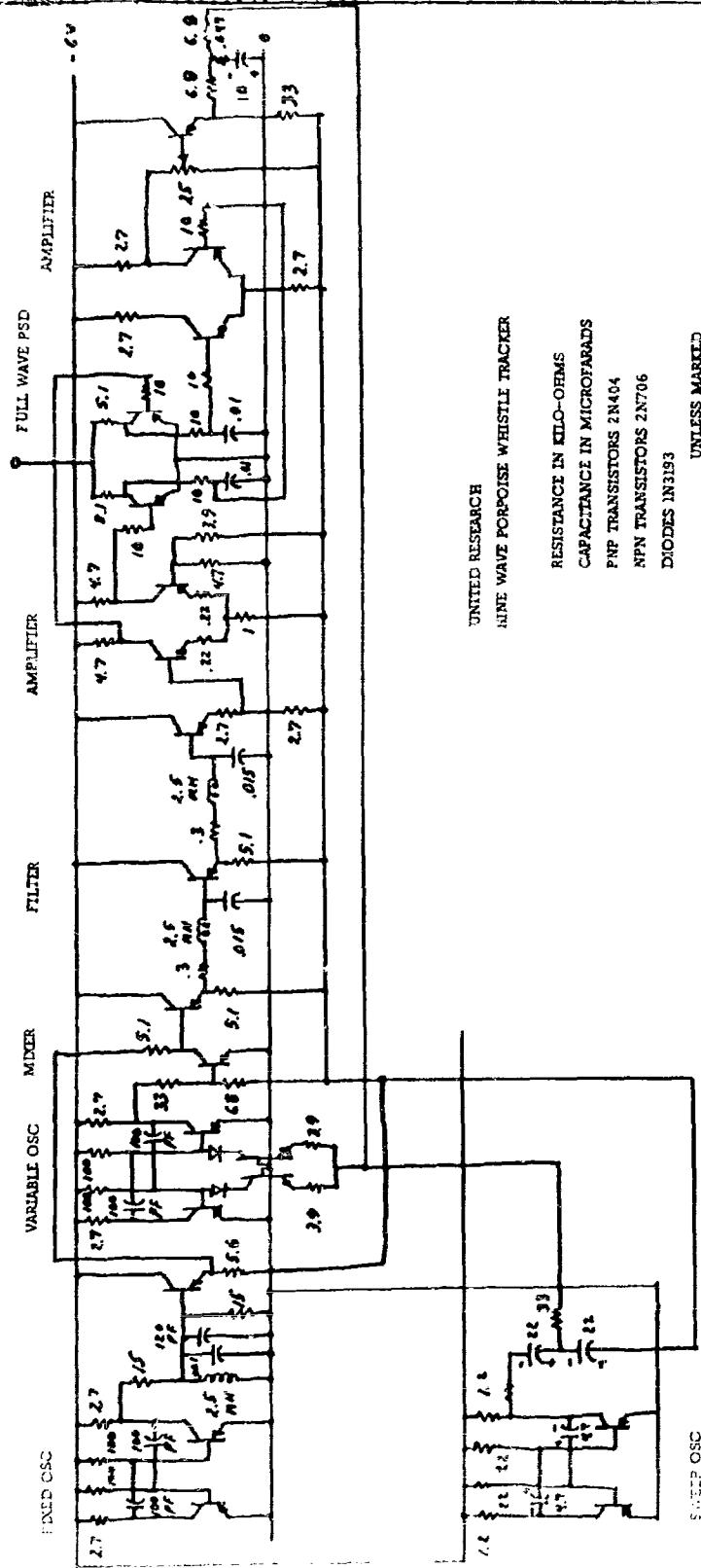
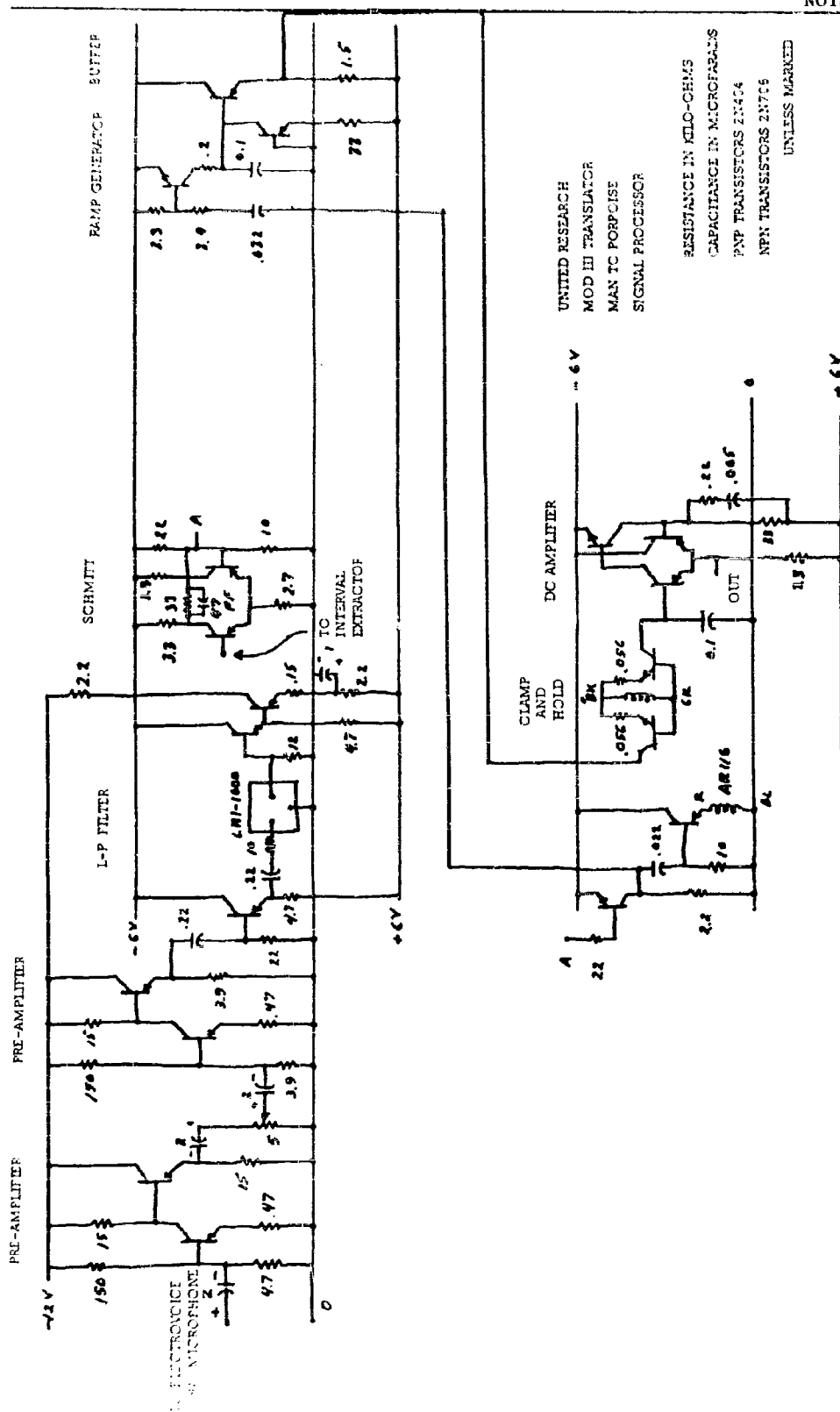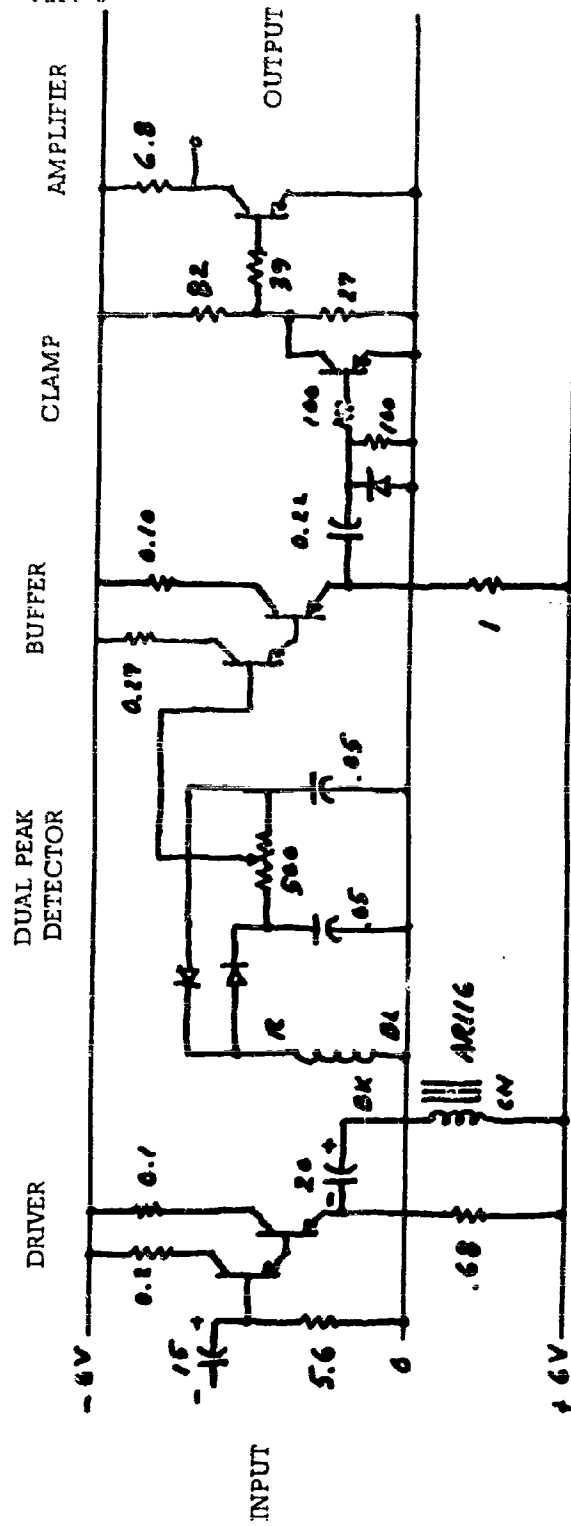Fig. 4-11   Circuit Diagram of Sine Wave Porpoise Whistle Tracker

Fig. 4-12   Circuit Diagram of Mod III Man-Porpoise Signal Processor

Fig. 4-13   Circuit Diagram of Mod III Man-Porpoise Voice Interval Extractor
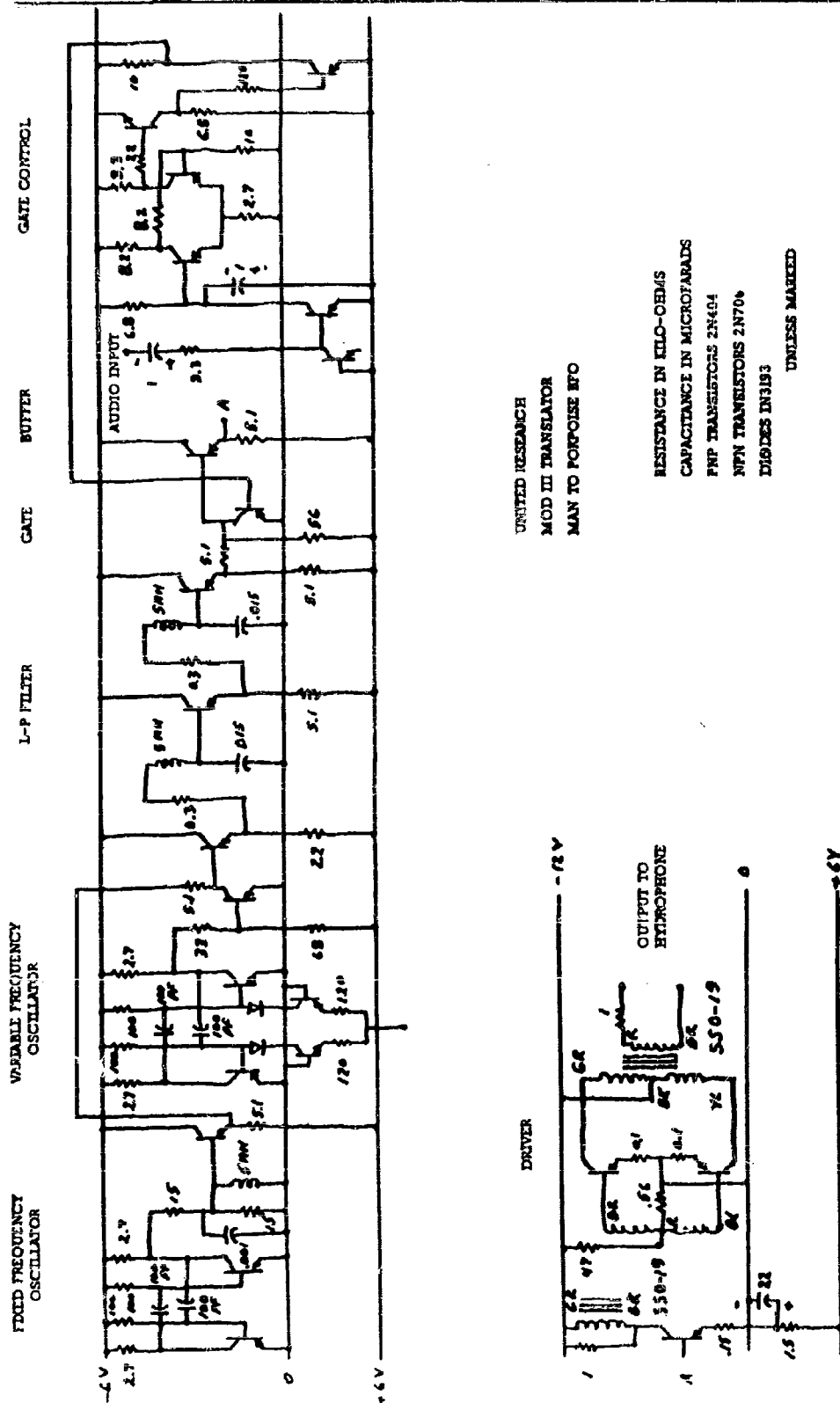
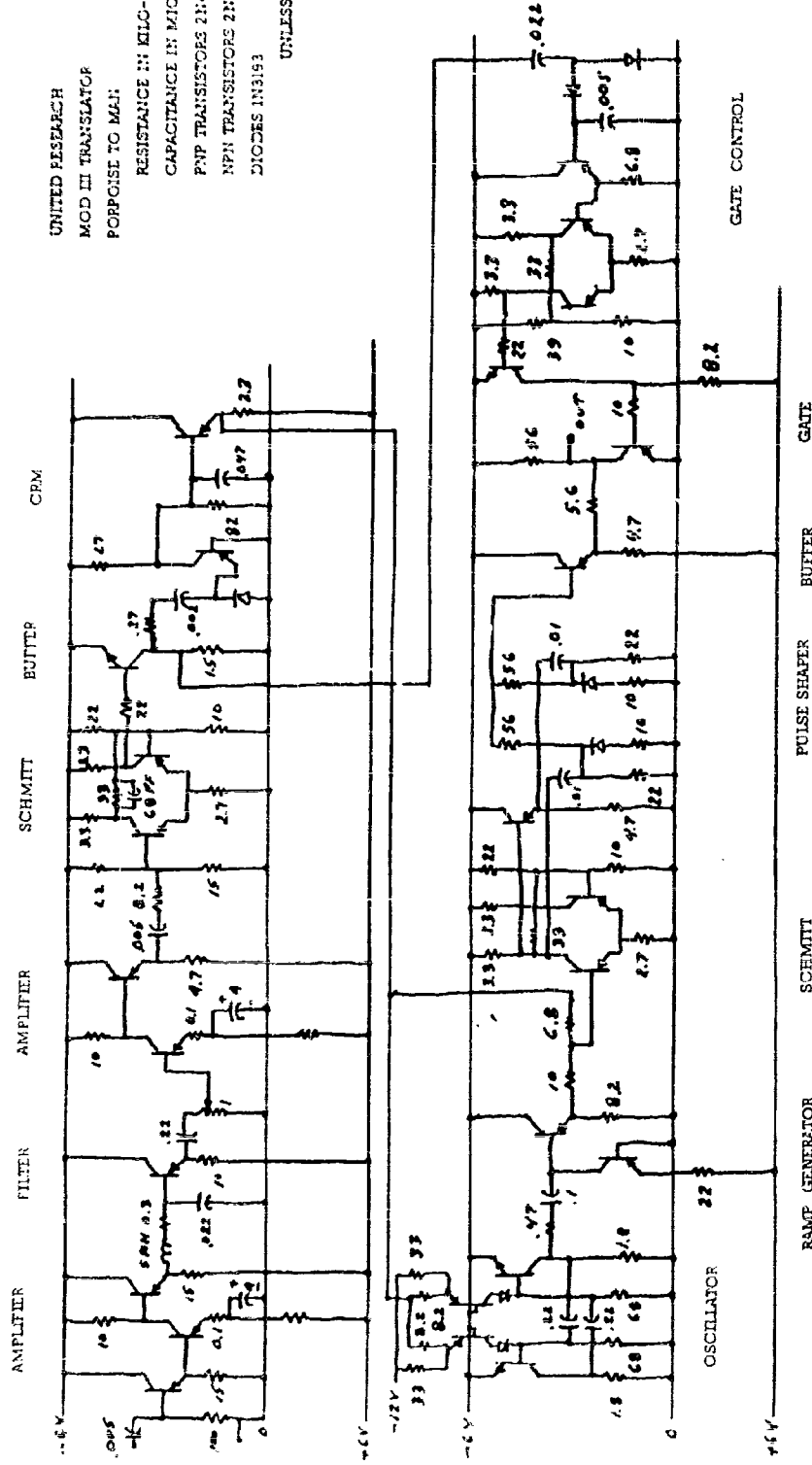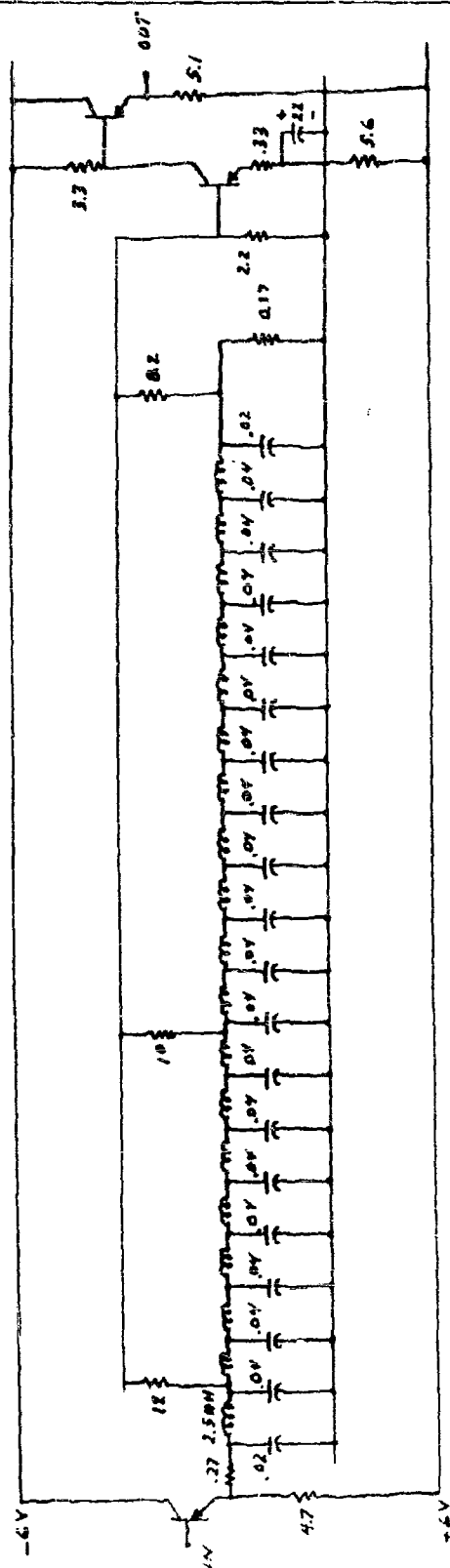Fig. 4-14 Circuit Diagram of Mod III Man-Porpoise BFO

Fig. 4-15   Circuit Diagram of Mod III Porpoise-Man Translator

UNITED RESEARCH

MOD III TRANSLATOR

PORPOISE TO MAN

DELAY LINE PROCESSOR

RESISTANCE IN KILO—OHMS

CAPACITANCE IN MICROFARADS

TRANSISTORS 2N404

Fig. 4-16   Circuit Diagram of Mod III Porpoise–Man Delay Line Processor

REFERENCES

2.1   von Bekesy, Georg, "Experiments in Hearing," McGraw-Hill, 1960.

2.2   Licklider, J.C.R., "Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech," Journal of the Acoustical Society of America, Vol. 20, No. 1, January 1948.

3.1   Batteau, D.W. "Characteristics of Human Localization of Sound," Final Report, May 1961, prepared under Contract No. N123-(60530) 23545A, for U.S. Naval Ordnance Test Station, China Lake, California, by United Research Incorporated, Cambridge, Massachusetts.

3.2   Batteau, D.W., Plante, R.L., "The Mechanism of Human Localization of Sounds with Applications in Remote Environments," Final Report, March 1962, prepared under Contract No. N123-(60530) 27872A for U.S. Naval Ordnance Test Station, China Lake, California, by United Research Incorporated, Cambridge, Massachusetts.

3.3   Batteau, D.W., "Localization of Sound: A New Theory of Human Audition," Final Report, September 1963, prepared under Contract No. N123-(60530) 30283A, for U.S. Naval Ordnance Test Station, China Lake, California, by United Research Incorporated, Cambridge, Massachusetts.

3.4   Batteau, D.W., Plante, R.L., "Further Developments in Sound Localization," Final Report, December 1963, prepared under Contract No. N123-(60530) 32279A, for U.S. Naval Ordnance Test Station, China Lake, California, by United Research Incorporated, Cambridge, Massachusetts.

3.5   Bruel, Per V., Frederiksen, Erling, Rasmussen, Gunnar, "Artificial Ears for the Calibration of Earphones of the External Type", Bruel and Kjaer Technical Review, No. 4, 1961.

3.6   Sessler, G.M., "Electrostatic Microphones with Electret Foil", Journal of the Acoustical Society of America, September 1963.