THE ANALYSIS OF COMPLEX SOUNDS

BY COCHLEAR PATTERNS

by

William Fleming Caldwell

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

1962
THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

I hereby recommend that this dissertation prepared under my direction by William Fleming Caldwell, entitled The Analysis of Complex Sounds by Cochlear Patterns, be accepted as fulfilling the dissertation requirement of the degree of Doctor of Philosophy.

Dissertation Director Date

After inspection of the dissertation, the following members of the Final Examination Committee concur in its approval and recommend its acceptance:

[Finger签字]

29 April '62

30 April '62

30 April '62

30 April '62

*This approval and acceptance is contingent on the candidate's adequate performance and defense of this dissertation at the final oral examination. The inclusion of this sheet bound into the library copy of the dissertation is evidence of satisfactory performance at the final examination.
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in The University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in their judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: [Signature]

APPROVAL BY DISSERTATION DIRECTOR

This dissertation has been approved on the date shown below:

[Signature]  
J. L. Stewart  
Professor of Electrical Engineering

Date: 16 April, 1962
ABSTRACT

An electrical analog of the human ear is developed to provide real-time cochlear patterns of subjective loudness along the basilar membrane. The outer and middle ear response is simulated in a functional manner. The equations of cochlear dynamics are developed and one-to-one relations between the physiological parameters of the cochlea and the circuit parameters of a lumped electrical analog are established. The necessary parameters of the cochlea are evaluated from experimental data. The mechanism of the sensory elements of the ear is functionally simulated by loudness converters in the analog. The resultant spatial pattern of loudness along the cochlea is converted into a time pattern for direct observation or for processing by subsequent devices.

It is hypothesized that cochlear pattern shapes are of principal significance in the recognition of sound. It is hypothesized that the cochlea performs a partial analysis of the sound and that the higher analysis centers of the central nervous system perform additional analyses that lead to the recognition process. The concepts and processes of analysis and recognition of sustained sounds are developed. An experiment is described which demonstrates the similarity in recognition between the human and the analog. The use of cochlear pattern recognition as the basis for a communication system is briefly discussed.
ACKNOWLEDGMENT

The writer wishes to acknowledge and express his gratitude for the guidance and encouragement of his dissertation director, Dr. J. L. Stewart. Acknowledgment is also made of the contributions to this research by his colleagues, Mr. E. Glaesser, Mr. K. D. Dimmick, and Mr. R. M. Brooks. Last, acknowledgment must be made of the constant encouragement of his wife, Joan Lippincott Caldwell, and his parents, Mr. and Mrs. L. C. Caldwell.

This research was in part supported by the Aerospace Medical Laboratory, Aeronautical Systems Division, Wright-Patterson Air Force Base, under contract No. AF 33(657)-8016.
# TABLE OF CONTENTS

## CHAPTER

1 **INTRODUCTION OF THE PROBLEM**
- 1.1 Introduction 1
- 1.2 The Problem Defined 4
- 1.3 Methods and Organization 4
- 1.4 Literature Survey 6

2 **THEORY OF THE ELECTRONIC ANALOG EAR**
- 2.1 Introduction 11
- 2.2 Structure of the Ear 12
  - 2.2.1 General structure 12
  - 2.2.2 The cochlea 20
- 2.3 Functional Realization of the Outer and Middle Ear 24
- 2.4 Equations for Motion in the Cochlea 27
  - 2.4.1 Model of the cochlea 27
  - 2.4.2 Equations of continuity 30
  - 2.4.3 Equations of fluid motion 31
  - 2.4.4 Equation of cochlear duct motion 32
  - 2.4.5 Combined equations of motion 33
- 2.5 Electrical Analog of the Cochlea 36
- 2.6 Sensory Loudness Conversion 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>PARAMETERS OF THE ANALOG COCHLEA</td>
<td>45</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>45</td>
</tr>
<tr>
<td>3.2</td>
<td>Dynamic Mass</td>
<td>48</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Energy equivalent mass</td>
<td>48</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Loading mass</td>
<td>52</td>
</tr>
<tr>
<td>3.3</td>
<td>Dynamic Friction</td>
<td>57</td>
</tr>
<tr>
<td>3.4</td>
<td>Dynamic Stiffness</td>
<td>60</td>
</tr>
<tr>
<td>3.5</td>
<td>Equivalent Friction of the Perilymph</td>
<td>62</td>
</tr>
<tr>
<td>4</td>
<td>DESIGN OF THE ELECTRONIC ANALOG EAR</td>
<td>65</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>65</td>
</tr>
<tr>
<td>4.2</td>
<td>Design of the Analog Cochlea</td>
<td>68</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Selection of the number of RLC sections</td>
<td>68</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Impedance scaling</td>
<td>69</td>
</tr>
<tr>
<td>4.2.3</td>
<td>The analog cochlea</td>
<td>70</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Readout of velocity and displacement</td>
<td>76</td>
</tr>
<tr>
<td>4.3</td>
<td>Design of the Loudness Converters</td>
<td>78</td>
</tr>
<tr>
<td>4.4</td>
<td>The Input Amplifier</td>
<td>81</td>
</tr>
<tr>
<td>4.5</td>
<td>Sampling and the Pattern Readout</td>
<td>86</td>
</tr>
<tr>
<td>4.6</td>
<td>Results from the Analog Ear</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td>THEORY OF COCHLEAR PATTERN ANALYSIS</td>
<td>92</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>92</td>
</tr>
<tr>
<td>5.2</td>
<td>Concepts of Pattern Theory</td>
<td>95</td>
</tr>
<tr>
<td>5.3</td>
<td>Concepts of Pattern Analysis and Recognition</td>
<td>96</td>
</tr>
</tbody>
</table>
5.4 Cross-correlation as a Recognition Process

5.4.1 Cauchy-Schwarz Inequality
5.4.2 Cross-correlation function
5.4.3 Mean square error

5.5 Absolute Difference as a Recognition Process

5.6 Recognition of Cochlear Patterns

6 COCHLEAR PATTERN ANALYSIS OF EIGHT VOWELS

6.1 Introduction

6.2 Eight Selected Sustained Vowels

6.2.1 The selected vowels
6.2.2 Human confusion matrices

6.3 Vowel Patterns From the Analog Ear

6.3.1 Introduction
6.3.2 Quantization and Sampling
6.3.3 Vowel pattern data

6.4 Cross-correlation of Eight Selected Vowels

6.4.1 Correlation matrices
6.4.2 Recognition matrices and recognition scatter diagram

6.5 Absolute Difference Analysis of Eight Selected Vowels

6.6 Vowel Pattern Similarity Among Speakers

6.7 Discussion of Results

7 SUMMARY AND RECOMMENDATION

7.1 Introduction

7.2 Discussion of the Research
7.3 Conclusions

7.4 A Communication System Using Recognition

7.5 Recommendations for Further Research
CHAPTER 1
INTRODUCTION TO THE PROBLEM

1.1 Introduction

The human ear is a remarkable device with which the human is able to detect and at least partially analyze sounds. The ear in conjunction with the central nervous system achieves recognition of complex sounds having a variety of combinations of frequencies and timing arrangements. This dissertation is concerned with an investigation of the human hearing and sound recognition system from the viewpoint of the concepts and techniques of analog modeling and of communication and control systems. The viewpoint is similar in nature to that of Cybernetics, which was defined by Wiener as "the entire field of control and communication theory, whether in the machine or in the animal." From this viewpoint the ear and central nervous system are parts of a communication system and thus will have those characteristics which are common among communication systems. By determining some of these characteristics and suggesting mechanisms by which some of the processes of a communication system may

---

be carried out by the human it will be possible to extend the general knowledge about the ear and hearing.

The sound analyzing capabilities of the ear were first discussed by Helmholtz in 1877 (ref. 2). In a classic paper, he discussed the mechanism of the cochlea, using as a model a series of coupled tuned strings which were excited when the proper resonant frequency was impressed on the input. Since that time, and particularly in the past half century, a number of studies have been devoted to an explanation of the operation of the ear, and especially of the cochlea. In most of these studies it was recognized that the key to the solution of the problem of sound recognition by the brain lay in understanding the mechanism of the cochlea and providing a proper interpretation of sounds in terms of the excitation of the sensory cells of the cochlea.

Another school of thought of interest has developed in conjunction with recent studies in voice communication systems. The principal effort of this school has been to develop communication systems with greater bandwidth efficiency, while retaining the ease of information encoding (speaking) and decoding (listening) of the human operator. The principal practical results of this work have been variations of the Vocoder principle. Advancing this technique, work has been done on systems wherein the spectrum pattern from the Vocoder is compared to a set of stored patterns and only the storage address or classification of the pattern is transmitted, to be reconstituted at the receiver from a local pattern storage.
Other speech processing systems have been hypothesized and investigated which describe speech in terms of the analog parameters of the speaking system. These, too, were intended to transmit in digital form a classification necessary to produce a particular sound. The majority of speech processing systems that have been studied in this manner may be described as spectral analysis schemes with various descriptive readouts. It is rather significant that at this point there were few attempts to establish a general connection between the study of the processing capability of the ear and central nervous system, and the study of speech processing for bandwidth reduction systems. This was probably due to ignorance, except in a most general way, of the operation of the cochlea.

Stewart, while engaged in a study of the loudness characteristics of the ear, proposed a "pattern theory" of sound recognition based on the pattern of loudness along the cochlea due to motion of the basilar membrane of the cochlea (ref. 3). This theory is more general than the place and volley theories, and it simplifies the explanation of a number of psychoacoustical phenomena. It is also, happily, more easily fitted into a mathematical model, thus having the advantage that results may be extrapolated from the theory, rather than only interpolated as in the case of semantic models.

The research described in this dissertation is in support of pattern concepts of sound recognition. And while principally concerned with expanding the knowledge of the ear and human hearing system, it will be
quite obvious that there are also applications to speech communications systems, when both are viewed from the common bond of sound recognition.

1.2 The Problem Defined

The fundamental hypothesis underlying this dissertation is that sounds are recognized by virtue of the patterns of motion created in the sensory structure of the cochlea. This ascribes to the cochlea the characteristics of a kind of spectrum analyzer, which in conjunction with the higher analysis centers of the central nervous system is capable of distinguishing members of an ensemble of patterns, which are derived from a like ensemble of distinguishable sounds. Therefore, an additional hypothesis may be stated: distinguishable sounds have distinguishable cochlear patterns.

It may be stated that the problem of this dissertation is: To study the analysis techniques necessary to characterize the process of sound recognition, and to demonstrate that distinguishable sounds have distinguishable cochlear patterns.

1.3 Methods and Organization

The mathematical representation for the sound-to-pattern transformation is quite complex. It is thus very advantageous to construct an electrical analog of the ear so that the patterns resulting from various classes of sounds may be conveniently measured and studied. Therefore,
it is first necessary to define, design, and build an electrical analog of the ear from which the spatial pattern of loudness along the cochlea may be displayed for study.

The development of an analog of the human ear is a particularly difficult task in view of the small amount of factual information that exists on the subject. This difficulty is amplified by the number and variety of "Theories of Hearing" which have been proposed; some with very involved reasoning for neglecting this or that parameter in the analysis of the action of the cochlea. Because an oversimplified theory can rarely describe more than a fraction of the observed phenomena, it is very desirable that the analog bear as close a one-to-one relation to the ear as is possible.

The next problem is to study the methods applicable to the analysis of the patterns of loudness from the analog. This may be done from the viewpoint of recognition; that is, methods of analysis should yield recognition similar to the human. Of particular interest at this point will be the establishment of an error criterion for use in recognition. While it is possible to suggest several criteria, it is not possible at this point to determine the one used by the human.

To further investigate the characteristics of pattern recognition it is necessary to compare human confusion in a recognition situation to the error expected from analysis techniques for similar patterns. For this purpose sustained vowels will be used, primarily because they are basic to
speech recognition. The eight vowels selected for study are easily produced, have good spectral spacing, and there is a great deal of information on them.

As to general format, this dissertation will develop a mathematical model of the cochlea from which an electrical analog will be set up. The outer and middle ear transfer characteristic will be functionally realized in the analog cochlea driver amplifier. Patterns of loudness converted motion of the basilar membrane, photographed from the display oscilloscope, will be sampled and quantized for analysis. A study will be made of applicable techniques of analysis for recognition of cochlear patterns and these techniques will be applied to the study of patterns of eight sustained vowels. The results of this study will be compared to the results of recognition tests performed by human subjects. The application of the electronic analog ear and pattern recognition to speech communication systems will be considered and briefly discussed.

1.4 Literature Survey

The literature concerning the ear and hearing is extensive and rather formidable. The references discussed here are not exhaustive, but represent the principal sources used in the development of this dissertation.

2 A derivation and discussion of a one-to-one network representation of the outer and middle ear is presented in the Doctoral Dissertation of E. Glaesser, University of Arizona, 1962.
In the introduction to this chapter the paper of Helmholtz was mentioned. This paper formed the basis for further thought and acted as a stimulus for numerous researches—it is generally regarded as the first important paper pertaining to the mechanics of the cochlea. The fifty year period succeeding this work was marked by the development of speculative theories which attempted to explain observed phenomena in hearing with little experimental support. While these theories were often based on models, they were generally unsatisfactory due to lack of knowledge of the physical mechanics of the cochlea.

As experimental techniques became better it was possible to study the action of the cochlea in some detail. The work of G. von Békésy (ref. 4) was particularly fruitful in terms of replacing speculation with fact—the present study has been built to a large extent on his experimental work. As with Helmholtz's work, Békésy's work led to a number of papers by other researchers. These papers mainly concern plausible mechanisms and the subsequent evaluation of parameters and responses using Békésy's experimental findings. The principal theoretical work has come from Békésy, Zwislocki (refs. 5, 6), Ranke (ref. 7), and Fletcher (refs. 8, 9).

Along with the theoretical investigations there has been a constant effort devoted to building models of the cochlea. Békésy built hydraulic and mechanical models based on his experimental work. Peterson and Bogert (refs. 10, 11) of Bell Laboratories built an electrical lumped parameter model of the cochlea based principally on the work of Békésy and Zwislocki. Their
original analog of the cochlea contained no dissipative elements, thus standing waves existed along the analog cochlea. It is now generally believed that the cochlea has sufficiently low Q so that it supports primarily a traveling wave. A later modification of this analog included dissipative elements, but there has evidently been little follow up on this work.

Bauch (ref. 12) built an electrical analog of the cochlea based on the theoretical work of Zwislocki. As Zwislocki had hypothesized frequency localization was demonstrated even though the mass of the cochlear duct was neglected. Bauch also inserted the cochlear duct mass into his analog and observed, for sine wave excitation, that there was again localization, though the response characteristics were greatly altered. Up to this point the models acted essentially as computers for calculating the motion of the basilar membrane.

Stewart (ref. 13) hypothesized the mathematical form of the conversion of sound excitation into loudness, and in subsequent investigations was able to demonstrate that sensory detection could be adequately described in this way. This work opened the field of hearing to the application of the techniques of modern communications and control theory. He further suggested that an analog of the ear should be built to study the patterns of loudness along the basilar membrane, hypothesizing that these patterns were of principal significance in the action of the ear and its function in hearing, both from detection and recognition points of view. It was reasoned
that this analog should hear as near a one-to-one relationship to the ear as possible so that the analog would have wide applications in the investigations of the ear and hearing. Two papers dealing with the theoretical aspects of this analog were presented to the Second Annual Bionics Symposium: The first (ref. 14) was concerned with loudness conversion, the second (ref. 15) was concerned with the theoretical and practical aspects of the design of an analog ear.

The latter part of this dissertation is concerned with the analysis and recognition of cochlear patterning. The field of pattern recognition is presently of great interest, and marked progress has been made in the theoretical fundamentals and practical techniques. In the analog ear discussed here, sequential sampling along the analog cochlea converts the original spatial pattern of loudness into a waveform in time. The problem of recognizing such a waveform has been dealt with to a considerable extent in modern communication theory, primarily in the theory of matched filters and the theory of correlation detectors. It has been shown that these two theories lead to equivalent mathematical operations, thus techniques based on either may conveniently draw from both.

Beyond the sources of general tutorial value, Cherry (refs. 16, 17, 18) and others have discussed the human hearing system from the viewpoint of a communication system. Of the prototype communication systems that have been publicly discussed, several make use of recognition of speech
elements. The general character of three such systems are discussed by Smith (ref. 19), Hoger and Putzrath (ref. 20), and Williams (ref. 21). In addition, there are a number of papers published which hypothesize communication systems based on speech element recognition. Some caution is necessary in estimating their value as the validity of their fundamental assumptions is often difficult to prove.
CHAPTER 2
THEORY OF THE ELECTRONIC ANALOG EAR

2.1 Introduction

In this chapter the theory of the electronic analog ear will be developed. It is necessary that this analog have the correct response from its input, which would correspond to a signal from outside the ear, to each point along the cochlea. In this dissertation the outer and middle ear will not be represented by a network on a one-to-one basis, but rather will be represented by a network, located in the amplifier driving the analog cochlea, which has the correct transfer function response. This was made necessary by the fact that the outer and middle ear networks were not completely evaluated at the time the present work was done.

The first step in this development is a description of the physiological structure of the ear. From this description an idealized hydraulic-mechanical model for the cochlea is hypothesized; from this the mathematical model for motion in the cochlea may be set up. The cochlea is analog modeled by means of an RLC lumped parameter electrical network which simulates the idealized hydraulic-mechanical model on a one-to-one basis. This network is similar in nature to a highly nonuniform, dissipative transmission line.
The hypothesis is made that the nerves leaving the cochlea transmit information which is some function of the envelope of the motion of the cochlear duct. What motion must be determined by experimental comparison of the human and analog response. From the theoretical aspects of loudness, a loudness converter is developed to represent the gross neural output of a small section along the cochlea.

2.2 Structure of the Ear

2.2.1 General structure

The general structure of the ear is shown in cross-section in Figure 2.1. The ear may be divided into three parts, the external ear, the middle ear, and the internal ear. These are shown in an idealized form in Figure 2.2. The external ear consists of the pinna and the external auditory meatus, the latter a tube closed by the tympanic membrane (eardrum). The middle ear is made up of the tympanic membrane and the ossicular chain, consisting of the malleus, incus, and stapes. These latter elements are contained within the middle ear cavity. This air filled cavity in the temporal bone is normally closed, although it may be equalized to external pressure by the eustachian tube. Thus normally, in spite of the local environmental pressure, the tympanic membrane will not have a bias pressure due to a difference in the pressure inside and outside of the middle ear.

The tympanic membrane is a flexible structure which closes the external auditory meatus. Rigidly attached to this membrane is the malleus.
FIGURE 2.1. CROSS-SECTION OF THE EAR
Figure 2.2. Idealized cross-section of the ear.
The malleus is in turn attached to the incus, which is in turn attached to, and transmits motion to, the stapes. This set of mechanical levers act as a pressure transformer between the external and internal ear. This transformer action is quite important as the tympanic membrane is a pressure transducer operating in air, while the stapes, which drives the oval window of the cochlea, is a transducer operating in a liquid. Thus the tympanic membrane and the ossicles provide a more efficient impedance match between the two media, air and liquid.

The internal ear is located within the temporal bone and is made up of two connected parts, the semicircular canals and the cochlea. The semicircular canals serve to sense motion and gravitational orientation. They are not important in hearing and will not be discussed further. The cochlea is, however, very important and a great deal of this dissertation will be concerned with its investigation. The cochlea consists of a spiral cavity within the temporal bone. This cavity contains fluid, various membranes, and the sensory elements of sound detection. The cochlear cavity is coupled to the middle ear by two openings, or windows, in the bone. These openings, the oval window and the round window, are each closed by a flexible membrane which serves to isolate the middle and internal ear. The stapes of the ossicular chain is attached to the membrane of the oval window in such a manner that force on the stapes is transmitted as a pressure into the fluid facing the oval window, in a
manner not unlike that of the piston of a pump. The round window and its membrane simply serve as a pressure relief for the cochlea. This pressure relief is, however, very important and the cochlea cannot function without it. This may be quite simply shown later with reference to the electrical analog circuit of the cochlea.

The spiraled cavity of the cochlea is divided along nearly its entire length by the cochlear duct. This flexible partition separates the cochlea into two scalae, the scala vestibuli, which is bounded by the oval window, and the scala tympani, which is bounded by the round window. At the apical end of the cochlea there is an opening, the helicotrema, not covered by the cochlear duct and thus directly linking the two scalae. The helicotrema prevents a fixed pressure from existing between the two scalae and thus across the cochlear duct.

The cochlear duct is itself fluid filled and made up or bounded by two membranes, Reissner's membrane and the basilar membrane. It is in this latter membrane that the sensory structures of the ear are imbedded. It is the characteristics of the cochlear duct, and especially of the basilar membrane, that determine the frequency localization characteristics along the cochlea.

Although the physiology of the ear will be discussed later in further detail, it is now possible to trace the auditory signal through its path from a sound wave outside the head to a neural signal from the cochlea. Sound vibrations strike the head and pinna, traverse the external meatus, and
excite the tympanic membrane. Movement of the tympanic membrane causes motion of the stapes due to the mechanical linkage of the ossicles. When the stapes moves it induces fluid motion; this fluid motion causes a sequence of traveling waves to traverse the scalae and excite the cochlear partition from the basal to the apical end. This motion in turn excites the sensory receptors within the cochlear duct, and thus causes a neural output. As a result of the limited firing rate of the individual neurons and the profusion of sensory elements, the spatial signal is assumed to consist of a pattern of loudness which is functionally related to the envelope of the traveling wave.

The tympanic membrane is a light but stiff cone with flexible edges which closes the external meatus diagonally as shown in Figure 23. Its apical angle is about 135 degrees and its total area is about 85 mm\(^2\), about 55 mm\(^2\) of which is rigidly connected to the malleus. \(^1\) At low frequencies the membrane vibrates with the malleus as a stiff cone about an axis near one edge. At higher frequencies the membrane loses its stiffness and the motion of the malleus lags the motion of the adjacent membrane.

The ossicles are a set of mechanical levers between the tympanic membrane and the oval window which provide a force amplification of about 1:1.3. This system of levers and the manner in which they are

---

AXIS OF ROTATION

MALLEUS
INCUS
STAPEDIUS MUSCLE
STAPES

TENSOR TYPANIC MUSCLE

EUSTACHIAN TUBE

TYMPANIC MEMBRANE

FIGURE 2.3. THE OUTER EAR AND MIDDLE EAR
attached to the tympanic membrane and the bone of the middle ear cavity is shown in Figure 2.3. The malleus, besides being attached to the tympanic membrane, is attached to the bone of the middle ear cavity by the superior malleolar ligament and the lateral ligament. The posterior aspect of the malleus has an irregular surface upon which is fitted a corresponding surface of the incus. The malleus and incus are joined at a stiff joint, the incudo-malleolar joint, and move as a single unit. The incus is also supported by a superior and a lateral ligament. Under normal conditions the ossicles tend to rotate about an axis parallel to and in approximately the same position as the anterior ligament, which lies just above the upper edge of the tympanic membrane. The rather large heads of the malleus and incus serve as counterweights for the long processes, such that the center of gravity and the center of rotation coincide. It has been observed that cutting the lateral ligaments has little effect upon the mode of vibration of the ossicles.

Two muscles, the tensor tympani and the stapedius muscle, are associated with the action of the middle ear. The first is attached to the malleus and the latter to the posterior aspect of the stapes. These muscles are normally relaxed, but reflexively contract in response to strong sounds. Contraction of the tensor tympani slightly tenses the tympanic membrane.

\[2\]
Ibid., p. 115.
while contraction of the stapedial muscle results in a shift of the axis of rotation of the stapes footplate and the oval window, such that a different mode of vibration occurs. Thus the stapes has essentially two modes of vibration that depend upon the state of the stapedial muscle.

The effect of the middle ear muscles is to change the mode of operation of the ossicles and/or stiffen membranes as a result of relatively intense sounds. This constitutes a feedback system. Unlike the simple automatic gain control, however, feedback in this case affects the system parameters.

One of the phenomena in hearing which is quite interesting is the very large dynamic range of about 120 db between threshold and feeling. Without the augmentation of the feedback this range is estimated to be in excess of 50 db. The analog of the ear will represent this unaugmented or small signal case and was designed to have a dynamic range in excess of 60 db. It was estimated that a dynamic range of 45 db is adequate for normal speech.

2.2.2 The cochlea

The cochlea is a snail shaped cavity of two and three quarters turns lying within the temporal bone. This cavity has its largest area and radius of curvature at the basal end, both of which decrease approaching the apical end. It is divided into two canals, the scala vestibuli and the scala tympani, by the cochlear duct. This flexible partition has a roughly triangular cross section and extends from the basal end to the apical end of
the cochlea, where there is a small opening, the helicotrema. A sketch of
a cross-section photograph of the cochlea is shown in Figure 2.4.

The scala vestibuli and the scala tympani are filled with perilymph,
a fluid having a density and a coefficient of viscosity about the same as
that of water. The cochlear duct, shown in detail in Figure 2.5, contains
endolymph, a viscous fluid. The basilar membrane is a fibrous elastic
membrane extending from the bony shelf to the spiral ligament. Reissner's
membrane is thin and very flexible.

The Organ of Corti, lying on the basilar membrane, contains about
20,000 hair cells with each cell containing four or five hair-like elements.
One end of each of these "hairs" extends through the scala media to the
tectorial membrane and is embedded into that membrane. The tectorial
membrane is composed of a system of diagonal fibers and a jelly-like
material that yields to slow movements, but is resistant to quick move­
ments. It is hinged like the cover of a book at the limbus.

As previously discussed, motion of the stapes causes a motion of
the fluid in the scalae and thereby motion of the cochlear duct. Bekesy
observed for low frequencies that, at any given location along the cochlea,
the Organ of Corti, the basilar membrane, the tectorial membrane, and
Reissner's membrane move in phase with one another. For frequencies

---

Ibid., pp. 442-443.
FIGURE 2.4 CROSS-SECTION OF THE COCHLEA
FIGURE 2.5. CROSS-SECTION OF THE COCHLEAR DUCT
above about 3000 cycles per second, Reissner's membrane and the basilar membrane are not in phase.

Motion of the basilar membrane creates a shearing action between the tectorial membrane and the reticular lamina. The relatively stiff reticular lamina tends either to rock upon its supports about an axis at the attachment of the basilar membrane to the bony modiolus, or to rock longitudinally, depending upon the location of the vibration along the cochlea. This shearing motion bends the hair cells between the tectorial membrane and the Organ of Corti, thus causing pulses from the nerves which are innervated with the hair cells.

The sensory receptors of the cochlea serve to produce the signals sent to the analysis centers of the central nervous system. For the pattern theory, it is not necessary that the cochlear output be of an exactly known form, but that it be known to within some one-to-one transformation in signal space. It is assumed that, when neural pulses from any particular segment of the cochlea reach the analysis centers, these pulses occur in such profusion as to be adequately represented by a continuous waveform.

2.3 Functional Realization of the Outer and Middle Ear

As previously discussed, sufficient data was not available to specify directly all of the necessary parameters associated with the middle and outer ear. Thus the known transfer response was simulated on a functional rather than a one-to-one basis. This was accomplished by means of a tuned
amplifier which provides the correct pressure transfer characteristic. This characteristic is necessarily correct only within a scale factor, which may be accounted for separately.

The transfer characteristic magnitude from the entrance of the external auditory meatus to the tympanic membrane is given by Wiener and Ross (ref. 22). The transfer characteristic magnitude of the middle ear, from the external side of the tympanic membrane to the cochlear side of the oval window, is given by Fletcher (ref. 9). By multiplying these experimentally determined characteristics, the overall response for the combined outer and middle ear is obtained. The separate curves and their resultant product response are shown in Figure 2.6.

It was determined that a suitable mathematical approximation for the overall transfer response is

\[
\frac{P}{P_{\text{ext}}} = \frac{j\omega \frac{G}{T_1}}{\left[1 + jQ \left(\frac{\omega}{\omega_o} - \frac{\omega}{\omega_c}\right)\right] \left[j\omega + \frac{1}{T_1}\right] \left[j\omega + \frac{1}{T_2}\right]}
\]

(2.3.1)

where

\[
G = 32 \quad (2.3.2)
\]

\[
Q = 2.5 \quad (2.3.3)
\]

\[
\omega_o = (2\pi)(4400) \quad \text{angular frequency} \quad \omega = 2\pi f \quad (2.3.4)
\]

\[
\frac{1}{T_1} = (2\pi)(15,000) \quad (2.3.5)
\]

\[
\frac{1}{T_2} = (2\pi)(1000) \quad (2.3.6)
\]
The mathematical approximation, which characterizes the actual amplifier is also shown in Figure 2.6. It is evident that this is an adequate approximation to the empirical data over most of the frequency range.

2.4 Equations for Motion in the Cochlea

2.4.1 Model of the cochlea

The hypothesized model of the cochlea is shown in Figure 2.7. It is assumed to be an idealized small signal cochlea, in which nonlinearities are ignored. From this model the partial differential equations of continuity and of motion are written to describe pressure and velocity distributions in the cochlea. The basic assumptions are:

1. The perilymph is an incompressible fluid.
2. Fluid motion in the scala vestibuli and scala tympani is negligible in all directions except along the cochlear axis.
3. The cochlear duct consists of independent elements of length \( \Delta x \), each with mass, elasticity, and friction.
4. The cross-sectional areas of the scala vestibuli and the scala tympani are equal.
5. The distension of each element of the cochlear duct is parabolic.

Based on these assumptions, the dynamics of the cochlea are expressed by the following equations:

1. Equation of fluid motion in the two scalae.
2. Equation of fluid continuity in the two scalae.
FIGURE 2.7 THE IDEALIZED COCHLEA
3. Equation of motion of an element of the cochlear duct.

To proceed with the mathematical development it is necessary to define the following parameters:

1. $A_1(x)$ - area of the scala vestibuli

2. $A_2(x)$ - area of the scala tympani

3. $A(x)$ - area of the scalae under the assumption of equal areas

4. $x, y, z$ - coordinates for the right hand coordinate system shown in Figure 2.7

5. $u_x, u_y, u_z$ - velocity components in the $x, y, z$ coordinates respectively

6. $b(x)$ - width of the basilar membrane

7. $p(x,t)$ - sound pressure, a function of distance and time

8. $m(x)$ - equivalent dynamic mass per unit length of the cochlear duct

9. $f(x)$ - equivalent dynamic friction per unit length of the cochlear duct

10. $k(x)$ - equivalent dynamic stiffness per unit length of the cochlear duct

11. $f_p(x)$ - equivalent friction of the perilymph per unit volume of the scala

12. $F(x,t)$ - force, a function of distance and time
Subscripts \( x, y, \) and \( z \) refer to the coordinate indicated by the letter. Subscripts 1 and 2 refer to the scala vestibuli and the scala tympani respectively.

### 2.4.2 Equations of continuity

Under the assumption that the perilymph is incompressible, the equation of continuity for the scala vestibuli is:

\[
\frac{d}{dt} \left( \text{volume} \right) = 0 = u_{x1}(x,t) A_1(x) - u_{x1}(x + \Delta x, t) A_1(x + \Delta x) - \frac{2}{3} b(x) \Delta x u_z(x,t) \quad (2.4.1)
\]

The last term is the result of the assumed distention of the element representing the cochlear duct.

For \( \Delta x \) small, the velocity along the cochlea may be approximated as:

\[
u_{x1}(x + \Delta x, t) = u_{x1}(x,t) + \frac{\partial u_{x1}(x,t)}{\partial x} \Delta x \quad (2.4.2)
\]

Also for \( \Delta x \) small, the cross-section area of the scalae may be expressed as:

\[
A_1(x + \Delta x) = A_1(x) + \frac{\partial A_1(x)}{\partial x} \Delta x \quad (2.4.3)
\]

Combining these approximations with equation 2.4.1 the equation for the continuity in the scala vestibuli is:
and similarly, the equation for continuity in the scala tympani is:

\[ 0 = - \frac{\partial}{\partial x} \left[ u_{x2}(x, t) A_2(x) \right] + \frac{2}{3} b(x) u_z(x, t) \quad (2.4.5) \]

These equations for the two scalae may be combined to form a set of two equations which express the conditions of continuity in the cochlea.

\[
u_z(x, t) = - \frac{3}{2} \frac{1}{b(x)} \frac{\partial}{\partial x} \left[ u_{x1}(x, t) A_1(x) \right] \quad (2.4.6)
\]

\[
u_z(x, t) = \frac{3}{2} \frac{1}{b(x)} \frac{\partial}{\partial x} \left[ u_{x2}(x, t) A_2(x) \right] \quad (2.4.7)
\]

2.4.3 Equations of fluid motion

The equations of fluid motion express the change in force across a fluid element, and thus the dynamics of the element. To obtain these equations, it is assumed that the flow of fluid is negligible in all directions except along the cochlear scala. The force across a fluid element is:

\[
\Delta F_{x1}(x, t) = F_{x1}(x, t) - F_{x1}(x + \Delta x, t) =
\]

\[
\left[ f_p(x) \frac{\partial x_1(x, t)}{\partial t} + \rho_p \frac{\partial^2 x_1(x, t)}{\partial t^2} \right] A_1(x) \Delta x \quad (2.4.8)
\]

This may be expressed in terms of pressure per unit length, upon dividing by the area, and as a function of the fluid velocity. Thus equation 2.4.7
becomes
\[
\frac{\Delta F_{x1}(x,t)}{A_1(x)\Delta x} = \frac{\Delta P_1(x,t)}{\Delta x} = f_p(x) u_{x1}(x,t) + \rho_p \frac{\partial u_{x1}(x,t)}{\partial t} \tag{2.4.9}
\]

This equation becomes, for the scala vestibuli:
\[
\begin{align*}
\frac{\Delta P_1(x,t)}{\Delta x} &= \frac{f_p(x)}{A_1(x)} \left[ u_{x1}(x,t) A_1(x) \right] \\
&\quad + \frac{\rho_p}{A_1(x)} \frac{\partial}{\partial t} \left[ u_{x1}(x,t) A_1(x) \right] \tag{2.4.10}
\end{align*}
\]

and similarly for the scala tympani:
\[
\begin{align*}
\frac{\Delta P_2(x,t)}{\Delta x} &= \frac{f_p(x)}{A_2(x)} \left[ u_{x2}(x,t) A_2(x) \right] \\
&\quad + \frac{\rho_p}{A_2(x)} \frac{\partial}{\partial t} \left[ u_{x2}(x,t) A_2(x) \right] \tag{2.4.11}
\end{align*}
\]

2.4.4 Equation of cochlear duct motion

The equation of motion for the cochlear duct expresses the motion of the duct as a function of the pressure differential between the scalae. To obtain this equation, it is assumed that the duct consists of independent elements of length \(\Delta x\), each possessing mass, elasticity, and friction. Alternately expressed, this amounts to the assumption that coupling along the cochlear duct takes place through the medium of the fluid in the scalae,
and not by mechanical attachment. The force across an element along the cochlea may be expressed as:

\[ F_x(x,t) = F_1(x,t) - F_2(x,t) \]

\[ = m(x) \Delta x \frac{\partial u_z(x,t)}{\partial t} + f(x) \Delta x \ u_z(x,t) \]

\[ + K(x) \Delta x \int u_z(x,t) \ dt \quad (2.4.12) \]

This may be expressed in terms of pressure by dividing by the area, \( b(x) \Delta x \), of the element along the cochlea.

\[ P_z(x,t) = \frac{m(x)}{b(x)} \frac{\partial u_z(x,t)}{\partial t} + \frac{f(x)}{b(x)} \ u_z(x,t) \]

\[ + \frac{k(x)}{b(x)} \int u_z(x,t) \ dt \quad (2.4.13) \]

2.4.5 Combined equations of motion

To summarize the developments of this section, the three sets of equations that form a system of equations which express the dynamics of the cochlea are:

**equations of continuity**

\[ u_z(x,t) = - \frac{3}{2} \ \frac{1}{b(x)} \ \frac{\partial}{\partial x} \left[ \frac{u_x(x,t)}{A_1(x)} \right] \quad (2.4.6) \]

\[ u_z(x,t) = \frac{3}{2} \ \frac{1}{b(x)} \ \frac{\partial}{\partial x} \left[ \frac{u_x(x,t)}{A_2(x)} \right] \quad (2.4.7) \]
equations of fluid motion

\[ \frac{\Delta P_1(x,t)}{\Delta x} = \frac{f_p(x)}{A_1(x)} \left[ u_{x1}(x,t) A_1(x) \right] \]
\[ + \frac{\rho_p}{A_1(x)} \frac{\partial}{\partial t} \left[ u_{x1}(x,t) A_1(x) \right] \quad (2.4.10) \]

\[ \frac{\Delta P_2(x,t)}{\Delta x} = \frac{f_p(x)}{A_2(x)} \left[ u_{x2}(x,t) A_2(x) \right] \]
\[ + \frac{\rho_p}{A_2(x)} \frac{\partial}{\partial t} \left[ u_{x2}(x,t) A_2(x) \right] \quad (2.4.11) \]

equations of cochlear duct motion

\[ \frac{Z}{b(x)} \left( \frac{m(x)}{b(x)} \frac{\partial u_z(x,t)}{\partial t} + \frac{f(x)}{b(x)} u_z(x,t) \right) + \frac{k(x)}{b(x)} \int u_z(x,t) \, dt \quad (2.4.13) \]

At this point in the development it is convenient to review the equations of motion in terms of what they physically describe. The set of equations formed by 2.4.10 and 2.4.11 describe the longitudinal pressure change along the cochlear scala due to friction and inertial mass of the fluid. Equation 2.4.13 describes the motion of the cochlear duct as a result of the pressure difference between the two scalae. Equations 2.4.6 and 2.4.7 serve to connect the longitudinal and transverse motions.
To more easily obtain the electrical analog of the cochlea, it is convenient to put the combined equations of motion into a different form. First, as a lumped parameter electrical analog is to be used to simulate the distributed parameters of the cochlea, the equations of motion must be in terms of finite differences instead of differentials. Next, the equations should be in a form where the elements of the analog are obvious by comparing the minimum number of equations. To do this, equation 2.4.6 is rewritten in terms of finite differences in the length, $\Delta x$, as

$$u_z(x,t) = -\frac{3}{2} \frac{1}{b(x)} \frac{\Delta (u_{x1}(x,t) A_1(x))}{\Delta x}$$  (2.4.14)

Substituting this into the equation for motion of the cochlear duct, equation 2.4.13, yields

$$P_z(x,t) = -\frac{3}{2} \frac{m(x)}{b^2(x)} \frac{1}{\Delta x} \frac{\partial}{\partial t} \Delta (u_{x1}(x,t) A_1(x))$$

$$- \frac{3}{2} \frac{f(x)}{b^2(x)} \frac{1}{\Delta x} \Delta (u_{x1}(x,t) A_1(x))$$

$$- \frac{3}{2} \frac{k(x)}{b^2(x)} \frac{1}{\Delta x} \int \Delta (u_{x1}(x,t) A_1(x)) \, dt$$  (2.4.15)

Then the equations for fluid motion, equations 2.4.10 and 2.4.11 may be rewritten as

$$\Delta P_1(x,t) = \frac{f_p(x)}{A_1(x)} \frac{\Delta x}{\partial} \left[ u_{x1}(x,t) A_1(x) \right]$$

$$+ \frac{\rho_p \Delta x}{A_1(x)} \frac{\partial}{\partial t} \left[ u_{x1}(x,t) A_1(x) \right]$$  (2.4.16)
\[ p_z(x,t) = \frac{f_p(x) \Delta x}{A_2(x)} \left[ u_{x2}(x,t) \frac{A_2(x)}{A_2(x)} \right] + \frac{\rho_p \Delta x}{A_2(x)} \frac{\partial}{\partial t} \left[ u_{x2}(x,t) \frac{A_2(x)}{A_2(x)} \right] \quad (2.4.17) \]

It is these equations, 2.4.15, 2.4.16, and 2.4.17 that will be used to compare the hydraulic-mechanical and electrical networks and to evaluate the electrical analog elements.

2.5 Electrical Analog of the Cochlea

There are several forms that the electrical analog of the cochlea may have, depending on the relations chosen for the analog equations. In this case a particular set of analog relations were chosen for their convenience in both setting up and constructing the electrical analog. The general cochlear analog section is shown in Figure 2.8a. This section is to represent an element of length along the cochlea. The equations for this network, written in terms of the variables indicated on the figure are

\[ e_z(x,t) = -L_p(x) \frac{\partial i(x,t)}{\partial t} - R_p(x) i(x,t) \]
\[ -\frac{1}{C_p(x)} \int i(x,t) \, dt \quad (2.5.1) \]
FIGURE 2.8a ONE SECTION OF THE COCHLEAR ANALOG

FIGURE 2.8b ONE SECTION OF THE COCHLEAR ANALOG IN UNBALANCED FORM
equation of the top series branch

\[ \Delta e_1 (x,t) = R_{s1}(x) i_1 (x,t) + L_{s1}(x) \frac{\partial i_1(x,t)}{\partial t} \]  \hspace{1cm} (2.5.2)

equation of the bottom series branch

\[ \Delta e_2 (x,t) = R_{s2}(x) i_2 (x,t) + L_{s2}(x) \frac{\partial i_2(x,t)}{\partial t} \]  \hspace{1cm} (2.5.3)

The following equations may now be directly compared:

- equation 2.4.15  ---  equation 2.5.1
- equation 2.4.16  ---  equation 2.5.2
- equation 2.4.17  ---  equation 2.5.3

and where the analog relations are:

- pressure  \rightarrow  voltage
- velocity \times area  \rightarrow  current

To establish the desired relations between the hydraulic-mechanical model of the cochlea and the proposed electrical model, the scale factor \( K_s \) and the impedance level factor \( K_z \) are introduced. It is defined that the voltage is equal to \( K_s \) times the pressure and that the electrical analog impedance is equal \( K_z \) times the mechanical impedance level of the actual ear.

Thus the relations between the variables are given by

\[ e_z (x,t) = K_s p_z (x,t) \]  \hspace{1cm} (2.5.4)
\[
I_2 = \Delta i(x,t) = \frac{K_s}{K_z} \frac{2}{3} b(x) \Delta x u_z(x,t)
\]  
\hspace{2cm} (2.5.5)

\[
\Delta e_1(x,t) = K_s \Delta P_1(x,t)
\]  
\hspace{2cm} (2.5.6)

\[
\Delta e_2(x,t) = K_s \Delta P_2(x,t)
\]  
\hspace{2cm} (2.5.7)

\[
\Delta i_1(x,t) = \frac{K_s}{K_z} \Delta \left( u_{x1}(x,t) A_1(x) \right)
\]  
\hspace{2cm} (2.5.8)

\[
\Delta i_2(x,t) = \frac{K_s}{K_z} \Delta \left( u_{x2}(x,t) A_2(x) \right)
\]  
\hspace{2cm} (2.5.9)

From the comparison of the two sets of equations the analogous elements of the electrical analog, in terms of the cochlear model, are:

\[
R_{s1}(x) = K_z \frac{f_p(x)}{A_1(x)} \Delta x
\]  
\hspace{2cm} (2.5.10)

\[
R_{s2}(x) = K_z \frac{f_p(x)}{A_2(x)} \Delta x
\]  
\hspace{2cm} (2.5.11)

\[
L_{s1}(x) = K_z \frac{P_p}{A_1(x)} \Delta x
\]  
\hspace{2cm} (2.5.12)

\[
L_{s2}(x) = K_z \frac{P_p}{A_2(x)} \Delta x
\]  
\hspace{2cm} (2.5.13)

\[
R_p(x) = K_z \frac{3}{2} \frac{f(x)}{b^2(x)} \frac{1}{\Delta x}
\]  
\hspace{2cm} (2.5.14)

\[
L_p(x) = K_z \frac{3}{2} \frac{m(x)}{b^2(x)} \frac{1}{\Delta x}
\]  
\hspace{2cm} (2.5.15)

\[
C_p(x) = \frac{1}{K_z} \frac{2}{3} \frac{b^2(x)}{k(x)} \Delta x
\]  
\hspace{2cm} (2.5.16)
To form the unbalanced equivalent of the network it is only necessary to make the series impedance of any given section of the unbalanced network equal to the series impedance of the corresponding section of the balanced network. Thus the series elements of the unbalanced network of Figure 2.8b are

\[
R_s = R_{s1} + R_{s2} = K_z f_p \Delta x \left[ \frac{1}{A_1(x)} + \frac{1}{A_2(x)} \right] \tag{2.5.17}
\]

\[
L_s = L_{s1} + L_{s2} = K_z \rho_p \Delta x \left[ \frac{1}{A_1(x)} + \frac{1}{A_2(x)} \right] \tag{2.5.18}
\]

Under the assumption of equal areas for the scalae, these equations become

\[
R_s = 2K_z \frac{f_p \Delta x}{A(x)} \tag{2.5.19}
\]

\[
L_s = 2K_z \frac{\rho_p \Delta x}{A(x)} \tag{2.5.20}
\]

The parallel branch is unchanged.

The velocity and displacement of the basilar membrane are found from the analog by measuring, respectively, the voltage across the resistor or the capacitor of the parallel branch. The voltage across the resistor is

\[
e_R^{e_p} (x,t) = I_z(x,t) R_p(x) \tag{2.5.21}
\]
Substituting for $i_z$ from equation 2.5.5, and for $R_p$ from equation 2.5.14, the velocity of any given point of the basilar membrane is given by

$$u_z(x,t) = e_{R_p}(x,t) \left[ \frac{b(x)}{f(x)} \right] \frac{1}{K_s} \quad (2.5.22)$$

Similarly the relation between capacitor voltage and displacement of the basilar membrane is

$$z(x,t) = e_{C_p}(x,t) \left[ \frac{b(x)}{k(x)} \right] \frac{1}{K_s} \quad (2.5.23)$$

2.6 Sensory Loudness Conversion

With each analog section representing a length $x$ of the cochlea there is associated a loudness converter. In order to simulate the gross response of the human to sound the loudness converter function is analogous to the hair cells, nerve structures, and the higher auditory centers. It has been hypothesized by Stewart that the sensory conversion of the ear is similar in nature to the simple detectors encountered in communications systems (refs. 13, 23).

Let $f(t)$ be a general sound waveform which represents effective total stimulation of the ear. This total stimulus, of mean square value $F$, consists of two independent additive parts $s(t)$ and $n(t)$, with mean square values $S$ and $N$ respectively. Then the loudness function proposed by Stewart is

$$L \propto \left\langle |f(t)|^n \right\rangle \quad (2.6.1)$$
Where the angular brackets, \( \langle \) and \( \rangle \), indicate an averaging process in time. If \( s(t) \) and \( n(t) \) are Gaussian random variables this becomes

\[
L \propto (F)^{n/2} = (S + N)^{n/2}
\]  
(2.6.2)

Subsequent analysis using this function shows that it models the sensory conversion system of the ear quite closely, giving an explanation of masking and several other psycho-acoustic phenomena. It is possible to measure the power law exponent \( n \) by psycho-acoustic tests which use the relative loudness of noise and a sine wave in noise. These tests indicate that \( n \) is approximately equal to 1. Thus the loudness converters are approximately linear envelope detectors.

Also of importance is the averaging time determined by the characteristics of the human hearing system. Although disputed among investigators, it is estimated to be about 0.2 seconds (ref. 24). This is adequately approximated by the first order function

\[
\frac{1}{1 + sT}
\]

where \( T = 0.2 \) seconds

(2.6.3)

The loudness converter may be represented by the clock diagram of Figure 2.9. The first block represents a linear gain and the second block represents the rectification process. Note that whether the rectification is half-wave or full-wave is immaterial as the difference is principally a gain change, for in this case the filtering time constant is sufficiently long that the ripple is not important. The third block represents an averaging
FIGURE 2.9 LOUDNESS CONVERTER BLOCK DIAGRAM
filter. It is very important that the loudness converter be properly implemented so that the response characteristics are correct. For instance, if a peak detector should be used, the phenomena of masking would not be evidenced, as masking exists only when the average of the sum and the sum of the average of the applied signals are unequal.
CHAPTER 3
PARAMETERS OF THE ANALOG COCHLEA

3.1 Introduction

To construct an analog of the cochlea it is necessary to evaluate the parameters of the cochlea. Three parameters may be obtained directly from experimental data.

1. \( \rho \) - Density of the perilymph
2. \( b(x) \) - Width of the basilar membrane, presented in graphical form in Figure 3.1.
3. \( A(x) \) - Area of the scala vestibuli and the scala tympani, presented in Figure 3.2.

The other parameters must be evaluated on the basis of theoretical calculations using experimental data taken from actual cochleae and from hydraulic models by Békésy and other investigators.

4. \( m(x) \) - Dynamic mass per unit length of the cochlear duct.
5. \( k(x) \) - Dynamic stiffness per unit length of the cochlear duct.
6. \( f(x) \) - Dynamic friction per unit length of the cochlear duct.
7. \( f_p(x) \) - Equivalent friction per unit volume of the perilymph.
FIGURE 3.1 WIDTH OF THE BASILAR MEMBRANE

\[ h(x) = \text{Basilar Membrane Width} \]

\[ x = \text{Distance from Oval Window} \]
Figure 3.2 Area of the Scalar
3.2 **Dynamic Mass**

3.2.1 **Energy equivalent mass**

The dynamic mass per unit length of the cochlear duct, \( m(x) \), consists of the sum of two components, energy equivalent mass per unit length, \( m_e(x) \), and loading mass per unit length, \( m_l(x) \). The energy equivalent mass per unit length, \( m_e(x) \), is defined as that mass per unit length which, when assigned to a point \( P \) at the center of the basilar membrane, possesses kinetic energy equal to that of the distributed mass of the entire cross-section of the cochlear duct under an assumed motion. The basic assumptions are:

1. Average density of the cochlear duct, \( \rho_D \), is unity.

2. Materials composing the cochlear duct are incompressible.

3. Membrane distension is parabolic.

4. The cochlear duct cross-section can be approximated by a combination of simple shapes, as indicated in Figure 3.3.

The method used to evaluate \( m_e(x) \) is similar to that used by Zwislocki.

To evaluate the energy equivalent mass, the kinetic energy of the duct cross-section is written as

\[
\text{Kinetic Energy} = T = T_a + T_b + T_c
\]

\[
T = \frac{1}{2} m_e(x) \left[ \frac{d}{dt} \left( \text{displacement of point } p \right) \right]^2
\]

\[
T = \frac{1}{2} m_e(x) v_b(x)^2
\]
FIGURE 3.9 ASSUMED CROSS-SECTION OF THE COCHLEA DUCT
To obtain expressions for $T_a$, $T_b$, and $T_c$, a parabolic distension of the membrane is assumed, and the magnitude of this distension is determined for various regions in the cross-section of the duct.

In the following equations the fact that the geometrical parameters of the duct cross-sections are functions of $x$, the distance along the cochlea, is not formally noted in order to simplify their writing. It is, however, implied that all of the geometrical parameters of the cochlear duct are functions of $x$.

In the section associated with $T_a$, referring to Figure 3.3, the displacement is given by

$$\gamma = 4A_R \frac{b - r}{R}$$

where the maximum distension, at $r = \frac{b}{2}$, is

$$\gamma_{\text{max}} = A_R b^2$$

In the section associated with $T_b$ and $T_c$ the displacement is

$$\gamma = 4A_R \left[ \frac{b_R}{B(z)} \right] y \left( B(z) - y \right)$$

where for the section associated with $T_b$

$$B(z) = b_c + z \left[ \frac{b_R - b_c}{h} \right]$$

and for the section associated with $T_c$

$$B(z) = b + z \left[ \frac{b_c - b}{h} \right]$$
The kinetic energies $T_a$, $T_b$, and $T_c$ are given as

$$T_a = \frac{\rho_D}{2} \int_0^{b_R} \int_0^r \left[ \frac{d\eta}{dt} \right]^2 d\phi dr$$

$$T_a = \frac{2}{15} D b^2 v_b^2 \phi$$

$$T_b = \frac{\rho_D}{2} \int_0^{h} \int_0^B \left[ \frac{d\eta}{dt} \right]^2 dy dz$$

$$T_b = \frac{4}{15} \rho_D \left[ \frac{b}{b_R - b_c} \right] b^2 v_b^2 \ln \frac{b_R}{b_c}$$

$$T_c = \frac{\rho_D}{2} \int_0^{h} \int_0^B \left[ \frac{d\eta}{dt} \right]^2 dy dz$$

$$T_c = \frac{4}{15} \rho_D \left[ \frac{h}{b_c - b} \right] b^2 v_b^2 \ln \frac{b_c}{b}$$

Therefore, the energy equivalent mass divided by $b^2(x)$ is

$$\frac{m_e(x)}{b^2(x)} = \frac{2}{v_b^2} \left[ T_a + T_b + T_c \right] \frac{1}{b^2(x)}$$
\[
\frac{m_e(x)}{b^2(x)} = \frac{4}{15} \rho_D \left[ \Theta + 2 \left( \frac{H}{b_R - b_C} \right) \ln \frac{b_R}{b_C} \right] + 2 \left( \frac{h}{b_C - b} \right) \ln \frac{b_C}{b} \tag{3.2.16}
\]

Where all geometrical parameters are functions of the distance along the cochlea.

To obtain numerical values for \( m_e(x) \) the relative dimensions of the cochlear duct must be known. These dimensions and the calculated energy equivalent mass divided by \( b^2(x) \) for the human cochlea shown in Figure 2.4 are listed in Table 3.1, and plotted in Figure 3.4.

### 3.2.2 Loading mass

There is a mass loading of the cochlear duct due to transverse motion of the fluid surrounding the duct. To evaluate the magnitude of this effect the equivalent loading mass is estimated from data obtained by von Békésy on a hydraulic model of the Cochlea.\(^1\) This model has a cochlear partition of the same volume elasticity as that of a human cochlea and a width equal to that of the basilar membrane. This two channel model was excited and it was observed that the location of the maximum displacement as a function of frequency was the same as for the human cochlea. Fluid was then removed from one channel and the maximum of displacement was again observed. It

---

\(^1\) von Békésy, pp. 437-441.
## TABLE 3.1

<table>
<thead>
<tr>
<th>$x$ (mm)</th>
<th>$\Theta(x)$</th>
<th>$b(x)$</th>
<th>$b_0(x)$</th>
<th>$b_R(x)$</th>
<th>$H(x)$</th>
<th>$h(x)$</th>
<th>$\frac{m_q(x)}{b^2(x)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>$44^\circ$</td>
<td>9</td>
<td>14.5</td>
<td>24</td>
<td>3</td>
<td>3</td>
<td>.398</td>
</tr>
<tr>
<td>16.5</td>
<td>$44^\circ$</td>
<td>13.5</td>
<td>21</td>
<td>25</td>
<td>3</td>
<td>3</td>
<td>.368</td>
</tr>
<tr>
<td>22.5</td>
<td>$30^\circ$</td>
<td>15</td>
<td>21</td>
<td>28</td>
<td>3</td>
<td>3</td>
<td>.296</td>
</tr>
<tr>
<td>27.5</td>
<td>$30^\circ$</td>
<td>22</td>
<td>30</td>
<td>32</td>
<td>3</td>
<td>3</td>
<td>.242</td>
</tr>
</tbody>
</table>
was noted that the point of maximum displacement could be maintained at a given position along the cochlea, before and after fluid was removed from one channel, by increasing the excitation frequency by approximately 20 percent.

The cochlear partition is assumed to consist of a series of independent elements, each possessing mass, elasticity, and friction. Thus for a given point along the cochlear duct, the frequency of maximum displacement \((\omega_m + \Delta \omega_m)\) with fluid in only one scala, and the frequency of maximum displacement \((\omega_m)\) with fluid in both scalae can be related as

\[
\omega_m = \alpha \sqrt{\frac{k}{m_e + 2m_1}}
\]

\[
\omega_m + \Delta \omega = \alpha \sqrt{\frac{k}{m_e + m_1}}
\]

where \(\alpha\) is a function of the damping. The damping has been shown to be approximately constant over the length of the cochlea.

The relation of the frequency at any given point with fluid in only one scala to the frequency with fluid in both scalae is calculated as

\[
\left[ \frac{\omega_m + \Delta \omega_m}{\omega_m} \right]^2 = 1.2 \left[ \frac{m_e + 2m_1}{m_e + m_1} \right]^2
\]

which is solved for \(m_1\) to determine the mass loading on one side of the

\[\text{Ibid.}, \ pp. \ 458-460.\]
In the actual ear the cochlear duct is bounded by the basilar membrane and Reissner's membrane. Below about 3000 cps the membranes vibrate in phase and thus the total loading mass is equal to the sum of the individual loading masses of the two membranes. It was assumed that this holds for the total length of the duct.

Equation 3.2.20 may be rewritten, on dividing by \( b^2(x) \), as

\[
\frac{m_e(x)}{b^2(x)} = 0.8 \frac{m_e(x)}{b^2(x)}
\]

where \( m_e(x) \) is a known quantity from section 3.2.1. It was estimated from the similarity of this problem and that of a square plate vibrating in a fluid that the elemental mass loading increases as to the square of the membrane width. Thus the total loading mass is given as

\[
\frac{m_l(x)}{b^2(x)} = 0.8 \frac{m_b(x)}{b^2(x)} \left[ 1 + \left( \frac{b_R(x)}{b(x)} \right)^2 \right]
\]

and the dynamic mass of the cochlear duct divided by \( b^2(x) \) becomes

\[
\frac{m(x)}{b^2(x)} = \frac{m_e(x)}{b^2(x)} \left\{ 1 + 0.8 \left[ 1 + \left( \frac{b_R(x)}{b(x)} \right)^2 \right] \right\}
\]

The approximation for the dynamic mass divided by \( b^2(x) \) as a function of the distance along the cochlea is shown in Figure 3.4.
3.3 **Dynamic Friction**

The friction per unit length of the cochlear duct is due to both the friction of the duct itself and the viscous friction loading by the fluid in the scalae. The total friction is obtained from data taken by Bekesy on the logarithmic decrement. 3

If the cochlea is excited by short pulses a damped oscillation is observed in the displacement of the basilar membrane. Bekesy measured these oscillations in fresh specimens and found that the ratio of amplitudes between successive oscillation peaks of the same polarity was about 4:1 to 6:1 over the entire length of the basilar membrane. Using an average ratio of 5:1, the corresponding logarithmic decrement is 1.6,

where the logarithmic decrement, \(D\), is defined as

\[
D(x) \approx \log_\beta \frac{M_1(x)}{M_2(x)} \tag{3.3.1}
\]

and \(M_1(x)\) and \(M_2(x)\) are the maximum magnitudes of successive deflections of the same polarity. The form of the motion is

\[
M(x,t) = M_0(x,t) e^{-\frac{\alpha t}{2}} \sin(\omega t + \phi) \tag{3.3.2}
\]

In terms of the parameters of the cochlea, the logarithmic decrement is

\[
D(x) = \frac{f(x)}{m(x)} \frac{1}{2f_r(x)} \tag{3.3.3}
\]

3 Ibid., pp. 458-460.
where \( f_r(x) \) is the damped resonant frequency of oscillation.

To find an expression for friction, equation 3.3.3 is rewritten as

\[
f(x) = 2D(x) \ f_r(x) \ m(x)
\]  

(3.3.4)

and as \( D = 1.6 \) this becomes

\[
f(x) = 3.2 \ m(x) \ f_r(x)
\]  

(3.3.5)

and the friction divided by \( b^2(x) \) is

\[
\frac{f(x)}{b^2(x)} = 3.2 \ \frac{m(x)}{b^2(x)} \ f_r(x)
\]  

(3.3.6)

For any given section of the cochlear duct the damped resonant frequency, \( f_r(x) \), and the undamped resonant frequency, \( f_o(x) \), are related as

\[
f_r(x) = \frac{f_o(x)}{\sqrt{1 + \left( \frac{D}{2\pi} \right)^2}}
\]  

(3.3.7)

For \( D = 1.6 \), it may be assumed that

\[
f_r(x) \approx f_o(x)
\]  

(3.3.8)

and equation 3.3.6 may be expressed as

\[
\frac{f(x)}{b^2(x)} = 3.2 \ \frac{m(x)}{b^2(x)} \ f_o(x)
\]  

(3.3.9)

In Figure 3.5 are four curves showing the localization of frequency along the basilar membrane. These are: 1) localization of maximum dis-
FIGURE 3.5 FREQUENCY LOCALIZATION ALONG THE COCHLEA
placement frequency, \( f_0(x) \), observed by Békésy, \(^4\) 2) localization of maximum velocity based on Békésy's data of curve 1 and the logarithmic decrement, 3) assumed localization of maximum velocity frequency, and 4) frequency localization from psychoacoustical tests. This latter curve does not necessarily imply localization of either a maximum of displacement or velocity and is provided only for purposes of comparison.

The friction divided by \( b^2(x) \) as obtained using the assumed velocity localization curve is shown in Figure 3.6.

### 3.4 Dynamic Stiffness

As it is assumed that the fluid of the scalae is incompressible the stiffness per unit length of the cochlear duct is due solely to the elasticity of the duct itself. To estimate the equivalent stiffness, the previously calculated mass and the location of the maximum of velocity along the cochlear duct as a function of frequency are employed. The expression for stiffness is

\[
k(x) = \omega_0^2(x) \ m(x)
\]

and the stiffness divided by \( b^2(x) \) is

\[
\frac{k(x)}{b^2(x)} = \omega_0^2(x) \ \frac{m(x)}{b^2(x)}
\]

\(^4\) Ibid., pp. 440-443.

\(^5\) Ibid., p. 442.
This function is plotted in Figure 3.7. The curve of Figure 3.7 actually agrees fairly well with data taken by Békésy using a fine hair probe with a known buckling force.  

3.5 **Equivalent Friction of the Perilymph**

To evaluate the equivalent friction of the perilymph it is assumed that each scala consists of a series of elementary tube sections of length Δx. Each of these cylindrical sections has constant cross-sectional area, equal to the area of the scala at that point. The method is that used by Zwislocki (ref. 5).

For any one section of constant cross-sectional area, the velocity \( u \) as a function of the radius \( r \) is given by

\[
 u(x, r) = \frac{p_0 - p_1}{4 \mu \Delta x} \left( R^2 - r^2 \right) \quad (3.5.1)
\]

where \( p_0 \) and \( p_1 \) are the pressures on the two ends of an elementary tube section. Then the average velocity of flow through the tube is

\[
 u(x) = \frac{1}{\pi R^2} \int_0^R 2\pi r u(x, r) \, dr \quad (3.5.2)
\]

\[
 = \frac{1}{\pi R^2} \int_0^R 2\pi r \left( \frac{p_0 - p_1}{4 \mu \Delta x} \right) \left( R^2 - r^2 \right) \, dr \quad (3.5.3)
\]

\[
 = \frac{p_0 - p_1}{8 \mu \Delta x} R^2 \quad (3.5.4)
\]

---

6 Ibid., pp. 467-468.
Figure 3.7 Stiffness of the Cochlear Duct

$x$ - Distance from Oval Window - mm
The relation between the force across the elementary tube section and the volume velocity is

\[ \Delta \text{Force} = f_p(x) \cdot \text{volume velocity} \cdot \Delta x \tag{3.5.5} \]

and the friction per unit volume can be expressed as

\[ f_p(x) = \frac{\Delta \text{Force}}{\text{volume velocity}} = \frac{(p_o - p_e)A(x)}{\text{volume velocity} \cdot \Delta x} \tag{3.5.6} \]

\[ f_p(x) = \frac{(p_o - p_1)\pi R^2}{8 \mu \Delta x} \cdot \frac{\pi R^2}{\Delta x} = \frac{8\pi \mu}{A(x)} \quad \text{where} \quad A(x) = \pi R^2 \tag{3.5.7} \]

The coefficient of viscosity \( \mu \) for the perilymph is 0.02. The equivalent friction of the perilymph is closely approximated by

\[ f_p(x) = \frac{1}{2 A(x)} \tag{3.5.9} \]
4.1 **Introduction**

In this chapter it is necessary to translate the physical model defined by theoretical analysis into an electrical analog from which information concerning the human ear may be deduced. There remain a number of problems in setting up the proper analog that should be considered in the perspective of a consistent design philosophy for the over-all device. The central aim was to build a research tool with the widest possible use and versatility in the study of the ear and hearing.

The block diagram of an analog system of the ear is shown in Figure 4.1. The amplifiers raise the signal level of a microphone, tape recorder, or other device so that it is of a sufficient amplitude and correctly matched in response (including the outer and middle ear response) to drive the analog circuits representing the ear.

In view of the very large dynamic range of the ear, it is well to consider comparable practical limitations of the analog. The upper limit is determined by the power available to drive the cochlea and the capacity of the individual cochlear sections to absorb power. The lower limit is determined by the ambient noise level from various internal noise sources.
Figure 4.1 Block diagram of the analog ear
Considering average speech intensities, it was estimated that a 40 to 60 db dynamic range would be sufficient for the study of speech patterns. In addition, it was estimated that a 20 db signal to noise ratio would be desirable at the lower level. It is emphasized that this estimate includes the system only up to the outputs of the individual loudness converters; it does not consider the noise introduced from the sampling operation.

In the matter of accuracy of construction, valuable hints were obtained from human physiology. Humans differ among one another to a considerable extent and thus the over-all precision of the line need not be great. However, human tissue is characterized by smooth and continuous changes, not discontinuities, in its mechanical properties. Thus the artificial cochlea was built to assure regular and uniform changes in element values from one section to the next.

The information desired from the analog cochlea is the pattern of motion of the basilar membrane; this pattern was presumed to be at least functionally related to the neural waveform as received by the auditory cortex. Motion is a rather inclusive word—is the important variable displacement or velocity, or some combination of the two? At this point the answer was unknown, therefore, provision was made to permit measurement of either of these quantities along the analog cochlea. In this way, test data from the analog cochlea, compared to similar data from the actual ear, was used to evaluate the contributions of displacement and velocity to waveform patterns in the auditory cortex.
In the analog of the cochlea, signals representing displacement or velocity of points along the basilar membrane are in the form of DC outputs. The spatial array of these outputs along the cochlea provides the desired pattern. A simple mechanical commutator was used, in which each single output is momentarily sequentially connected to the commutator output. Display of the patterns on an oscilloscope was then elementary.

4.2 Design of the Analog Cochlea

4.2.1 Selection of the number of RLC sections

The key to the design of an analog ear is the cochlea; it is the critical item in the selection of an appropriate impedance and signal input level. The first design problem was that of selecting the number of RLC sections to be used to represent the cochlea. There is no unique answer to this question. The ultimate answer must be based on the information requirements for recognition by subsequent analysis devices. Thus the selection of the number of RLC sections of the analog cochlea was relegated to a "best engineering guess." There were, however, several references available that made this guess less difficult to make.

Previous electrical analogs of the cochlea had been constructed from which data was taken by point to point measurements. Though the published information from these analogs is only for sine wave excitation, it serves as an indication of the number of segments into which the cochlea need be divided. Bogert and Peterson (ref. 10) used 165 sections and Bauch (ref. 12)
used 65. In both cases only the cochlea was modeled. After studying the rather broad patterns obtained from these analygs, it was evident that they contained more than an adequate number of sections to define the patterns of speech. Thus for convenience, the number of sections was set at 36; each section representing one millimeter of an idealized 36 millimeter cochlea. To make the "tee" representation consistent, the equivalent of one half millimeter was added to each end of the 35 millimeter cochlea used in plotting the graph of cochlear parameters.

It should be noted that the pattern concept does not require minute detail in the patterns as might other theories of perception and recognition.

4.2.2 Impedance scaling

The second design problem was impedance scaling from the basic analog equations. These equations are based on physical parameters and thus have the mechanical impedance level of the actual ear. The impedance scale factor $K_z$ was introduced in Section 2.5 to allow the electrical analog impedance level to be changed from that of the actual cochlea.

To select the practical value of $K_z$, it was noted that the elasticity, Figure 3.8, varied along the cochlea by about five orders of magnitude. In the practical line, the smallest value of capacity must be of sufficient size to avoid its being swamped by the stray distributed capacity of the circuit. Also, it is not desirable for the low frequency sections to have very large capacitors. By calculating the necessary range of capacity, and the related inductor size, it was possible to select a satisfactory compromise impedance scale factor.
Using this method of selecting a value of $K_z$, a convenient value of $1/40$ was selected to give

$$C_{p \text{ max}} = 6.66 \text{ microfarad}$$

$$C_{p \text{ min}} = 43.0 \text{ picofarad}$$

Note that this corresponds to a lowering of the impedance level by a factor of 40.

4.2.3 The analog cochlea

The electrical analog of the cochlea was developed in Section 2.5 as a series of unbalanced RLC "tee" sections. One typical "tee" section was shown in Figure 2.8b. The analog cochlea has a series structure, shown in Figure 4.2a, of 36 of these "tee" RLC sections. The elements of the individual "tee" sections may be computed using the equations of Chapter 2 and the appropriate parameter curves of Chapter 3.

To simplify the construction of the analog cochlea the series elements, $R_s$ and $L_s$, of adjacent RLC half-sections are added together so that between shunt branches there are two, instead of four elements. Thus, the analog cochlea will be of the form shown in Figure 4.2b.

The curves of Figures 4.3, 4.4, 4.5, and 4.6 present the information necessary for evaluating all the elements of the analog cochlea. Each curve is correct for $K_z$, the impedance scale factor, of $1/40$. To determine the shunt elements, read the curves of $R_p$, $L_p$, and $C_p$ at points where $x$, the
BASIC ANALOG LINE OF "T" SECTIONS

DERIVED ANALOG LINE

FIGURE 4.2  ANALOG OF THE COCHLEA
FIGURE 4-5 \( C_p \) - PARALLEL BRANCH CAPACITY

\( x \) - Distance from Oval Window - mm

CAPACITY - As Indicated

1000000
100000
10000
1000
100
10
1

MODEL
DATE

74
distance, is 0, 1, 2, ..., 35 millimeters. To determine the internal series elements, read the curves at a point midway between the points for the shunt elements. This takes into account the addition of adjacent half sections. The end series elements are found by taking one half of the values read at 0 and 35 millimeters.

4.2.4 Readout of velocity and displacement

It was previously stated as desirable that loudness converted velocity or displacement be read from the cochlea. To accomplish this it is convenient to measure the current (velocity) or the charge (displacement) in the shunt branches of the analog cochlea. The relations between these analog quantities are given by equations 2.5.22 and 2.5.23.

The most convenient way to measure the current in this series circuit is to read the voltage across $R_p$, as shown in Figure 4.7. By proper gain scaling the output of the loudness converter will correspond to the amplitude of the velocity of that section of the basilar membrane. As the total shunt resistance is known, the gain scaling is realized by tapping down on this resistor to form an attenuator.

To measure the charge in the shunt circuit it is convenient to read the voltage across the capacitor, as shown in Figure 4.7. Again to provide the proper gain scaling, an attenuator was devised across $C_p$. A bit of caution must be observed in the design of these attenuators, as in some sections the stray capacity shown as the dashed capacitors $C_1$ and $C_2$ will
FIGURE 4.7 ONE ANALOG COCHLEA RLC SECTION
cause the attenuator to be incorrect and frequency sensitive. To correct this, a capacitor $C_3$ is added to the attenuator circuit such that $(C_2 + C_3) R_{gc} = C_1 R_1$, which cancels the stray transmission zero with a transmission pole positioned by $C_3$. This effectively adds a capacitor attenuator in parallel with the resistor attenuator.

4.3 Design of the Loudness Converters

The theory of operation of the loudness converters was discussed in section 2.6, primarily in terms of the fundamental operations that are involved. In addition to these fundamental operations, it was convenient in the design of the analog ear for the loudness converters to have gain. The considerations in arriving at this decision and the values of gain were principally dynamic range, estimated input level, desired output level, and ease of construction. It was further decided that all of the conditions involved in the proper scaling of voltages into velocity and displacement could be met using two values of gain: a gain of 10 applies to the first 18 sections of the analog cochlea, and a gain of 320 applies for the last 18 sections of the cochlea. These gains are specified to be from the ac input signal to the dc output signal. The actual total gain requirements were calculated using equations 2.5.22 and 2.5.23, and are shown in Figure 4.8.

To implement the block diagram of Figure 2.9, the circuit of Figure 4.9 was designed. A double triode is used as a two stage RC coupled amplifier,
Figure 4.9 Loudness Converter

<table>
<thead>
<tr>
<th>LOUDNESS CONVERTER</th>
<th>GAIN</th>
<th>$R_c$</th>
<th>AMPLIFIER TUBE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIGH FREQUENCY - LOW GAIN</td>
<td>10</td>
<td>7.5K</td>
<td>12AU7A</td>
</tr>
<tr>
<td>LOW FREQUENCY - HIGH GAIN</td>
<td>320</td>
<td>1.8K</td>
<td>12AX7A</td>
</tr>
</tbody>
</table>
which may be made to have a gain of 10 or 320 by selecting the proper tube and cathode resistors indicated in Figure 2.9. This amplifier drives a cathode follower, which is capacity coupled to a rectifier circuit consisting of two 1N67A diodes and a 180K ohms resistor. The output voltage of the rectifier is then filtered by a single section of RC filtering with a time constant of about 50 milliseconds.

The characteristics of one of the amplifiers with a gain of 10 is shown in Figure 4.10. It is approximately a linear detector, as section 2.6 indicated it should be. Some waveform sensitivity exists, as a change in gain resulted when individual values of two non-harmonically related sinewaves were varied while the rms sum was kept constant. This sensitivity is, however, small.

In addition to the theoretical work on which this design for the loudness converters is based, it is believed that the effect of neural volleys can be taken into account by modification of the loudness converters, such that their response falls off below a frequency of about 1000 cps. This response can be accomplished by changing the 0.051 microfarad coupling capacitor between the amplifier stages to a 0.001 microfarad.

4.4 The Input Amplifier

The input amplifier amplifies the input from a variety of devices, taking into account the necessary response compensation, and presents this signal to the analog circuits that follow. The output level is sufficiently high so
that the output of the cochlea has an acceptable signal to noise ratio. As
the characteristics of the outer and middle ear were functionally realized,
the input amplifier also simulates this response.

The input pre-amplifier is a McIntosh C-8 with selectable input for
microphone, tape recorder, record player, and tuner. The output is avail­
able from a cathode follower at about a 2 volt level.

In section 2.3 it was stated that the outer and middle ear may be
represented by the transfer function of equation 2.3.1. This may be realized
by a series of three circuits, a lead network, a lag network, and a parallel
tuned network; as shown in Figure 4.11. Thus the transfer function, in
terms of the circuit elements, becomes

\[
\frac{P}{P_{\text{ext}}} = \frac{j\omega \frac{G}{T_1}}{1 + jQ\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right)}\left[j\omega + \frac{1}{T_1}\right]\left[j\omega + \frac{1}{T_2}\right]
\]  

(4.4.1)

Where the values of the parameters are

\[G = 32\]
\[\frac{1}{T_1} = \frac{1}{R_1C_1} = 2\pi (15,000)\]
\[\frac{1}{T_2} = \frac{1}{R_2C_2} = 2\pi (1000)\]
\[\omega_o = 2\pi (4400)\]

This circuit is realized in Figure 4.12.
FOR EACH BLOCK IDEAL VOLTAGE SOURCES AND INFINITE IMPEDANCE LOADS ARE ASSUMED

\[ \omega_0 = \frac{1}{\sqrt{L_3 C_3}} \]

\[ Q = \frac{R_3}{\omega_0 L_3} \]

**FIGURE 4.11** OUTER AND MIDDLE EAR FUNCTIONAL REALIZATION
INPUT PREAMPLIFIER AND POWER AMPLIFIER WITH OUTER AND MIDDLE EAR NETWORK

Figure 4.12 DRIVER AMPLIFIER
It was estimated that the maximum signal applied to the analog cochlea would be about 90 volts peak-to-peak. To obtain this voltage from an amplifier while maintaining low distortion, it was necessary to rather carefully select a tube and its operating conditions. Investigation revealed that a 4146 power tetrode tube connected as a triode would yield a very linear response over a wide swing in plate voltage. Thus it was used in the class A resistance-capacity coupled output amplifier shown in Figure 4.12. The output impedance of this amplifier is about 500 ohms. It was estimated that this was sufficiently low to approximate an ideal voltage source to the analog cochlea. The output circuit coupling capacity necessary to give a low coupling impedance is quite large, about 24 microfarads.

4.5 Sampling and the Pattern Readout

The spatial output along the cochlea is represented by the outputs of 36 loudness converters. For convenience the outputs from the loudness converters are sampled sequentially along the cochlea, thereby converting the spatial pattern into a time pattern. This time pattern is then available for display on the oscilloscope or for manipulation in other devices.

To arrive at an adequate sampling rate, several points were considered. First, it was desirable that the sampling rate be sufficiently high so that the patterns portray the distinguishable characteristics necessary for speech recognition. To accomplish this, it is necessary that the time required to sample a complete cochlear pattern be brief compared to duration
of the pattern, so that gross pattern changes do not occur during a single sampling period.

There are two clues to the evaluation of the time required for sampling. One is that the time constant at the output of a loudness converter is about 150 milliseconds. Thus the rate of change of the amplitude of the pattern at a point on the cochlea is limited.

The next clue comes from the study of the perception of time altered speech. Here, a switch (with equal on and off times) was employed to sample speech. The frequency of the sampling was varied and it was noted that when the frequency was above about 14 cycles per second or more, the speech was quite intelligible.

From this information it was estimated that the sampling along the cochlea should take place in under 25 milliseconds. After surveying the commutators available for sampling, a mercury jet unit with 100 segments and a rotation rate of 1200 rpm was selected. By placing the first 50 segments in parallel with the last 50, and thus obtaining two sweeps per revolution of the commutator, it was possible to sample the cochlea 40 times per second. Thus the 36 sections of the cochlea are sampled in 18 milliseconds.

This sampling rate is deemed adequate for the study of the sustained portion of the vowels and the continuant fricatives. Overall patterns of speech are also apparent, although some transitions in connected speech may be
observed. Rather detailed studies of the faster changing features of speech will be necessary before the validity of this sampling rate can be determined.

It was desirable that the patterns to be displayed on the oscilloscope have a reasonably smooth shape. But, since the commutator has an interval between each sample equal to a sample period, the output from the commutator was made up of a series of rectangular pulses. One way to smooth this output was to stretch each sample so that the interval between it and the next sample was filled. This could be accomplished by delaying each rectangular pulse for a period equal to the space between pulses, and then summing the original and the delayed pulse trains. This was done in the readout system of the analog ear by the insertion of a simple network after the commutator. The waveform was then additionally smoothed with a single RC filter to obtain a pattern that was quite acceptable.

In working with the analog ear it was thought convenient to have available a measure of the loudness of the signal as indicated by the pattern on the cochlea. Note that this is a subjective measure in the human and not easily related to the input signal. The hypothesis was made that the total subjective loudness of the human and the analog ear is made up of the sum of the loudness amplitudes at each point along the cochlea. Thus a loudness meter need only measure the area of the pattern. This is easily done as the spatial loudness patterns have been converted into repetitive
time waveforms which are only positive with respect to zero voltage. Thus a dc meter which reads the average component of a signal was connected directly to the output of the commutator or smoothing filter to read a measure of subjective loudness.

4.6 Results from the Analog Ear

The analog ear was assembled in a cabinet as shown in Figure 4.13, after which minor troubles were corrected and the unit calibrated for operation. On completion of this work a number of measurements were carried out in order to compare the analog to the human ear. The detail data on the analog ear as a network is presented by Glaesser (ref. 25). However, it may be briefly summarized as follows.

**The analog cochlea**

1. exhibits the correct input impedance
2. exhibits the correct transfer impedance
3. exhibits the correct frequency localization

**The analog ear**

1. exhibits the correct threshold curve—for loudness converted basilar membrane velocity
2. exhibits masking of high tones by lower tones

Patterns of loudness converted basilar membrane velocity and the frequency localization curve are shown in Figure 4.14.
FIGURE 4.13 THE ANALOG EAR
Loudness Patterns for Sine Wave Excitation of 100, 315, 1000, 3150, and 10,000 cps

CURVE OF FREQUENCY LOCALIZATION ALONG THE COCHLEA

FIGURE 4.14
5.1 Introduction

In the preceding chapters an electrical analog of the human ear has been developed for use in studying the ear and hearing. The following work will be devoted to theoretical aspects of recognition of cochlear patterns and to demonstration of the analytical characteristics of cochlear patterns.

It has been hypothesized that the sound-to-pattern transformation constitutes a partial analysis of the sound. It was further hypothesized that subsequent analysis of the pattern by the central nervous system provides recognition of the pattern and thereby of the sound.

To discuss the recognition of cochlear patterns it is necessary to describe the patterns themselves. There is evidence that the proper pattern to consider is that of loudness converted velocity, as its curve of loudness threshold in the analog corresponds most closely to that of the human. In addition, there is physiological reason to believe that the velocity and not the displacement of the hair cells yields neural output of the sensory structures. In view of this reasoning, loudness converted velocity patterns will be used in this study.
From the properties of the cochlea it is obvious that the pattern will have fixed length, equal to the length along which the sensory structures are distributed in the cochlear duct. It is also obvious that the end of the cochlear duct nearest the stapes is fixed, as it is connected to the walls of the scalae. The helicotrema end is open between the scalae and cannot support a pressure difference. Thus there may be a discontinuity in the pattern only at the helicotrema end. The polarity of the output is positive only in that it represents an absolute magnitude.

Therefore, the cochlear pattern is limited to a fixed spatial region and to only one polarity. No generality is lost in the analog by sequentially sampling along the cochlea, as the fixed spatial region is merely converted into a fixed temporal waveform determined by the sampling period. A typical cochlear pattern display is shown in Figure 5.1a and the representation to be used in the following theoretical development is shown in Figure 5.1b.

In the practical case of the analog ear there are 36 sample points along the cochlea, while there are 50 commutator segments per complete sample sweep. The first 10 segments are grounded, the next 36 segments sequentially sample the analog cochlea, and the last 4 segments are grounded. Thus the patterns must initially and finally be zero.

It follows that the general cochlear pattern of spatial period $L$ may occupy within bandwidth limitations the indicated region in Figure 5.1b, and be described as a function $f(x) \equiv 0$, where $f(x \leq 0) = f(x+L) = 0$. 
FIGURE 5.1a

Cochlear Pattern Period

FIGURE 5.1b
5.2 Concepts of Pattern Theory

The pattern theory constitutes a general theory of sensory stimulus recognition, which applies when two or more different measures of sensory stimulation are available for analysis by the central nervous system. For example, the sensor for smell (olfaction) has numerous olfactory receptors with various response characteristics. The pattern of stimulation of these receptors is thought to define the impression of the odor on the central nervous system. Likewise, for the sense of taste (gustation) there are various taste buds on the tongue, each having different response characteristics. It is hypothesized that the pattern of stimulation of these taste receptors defines a complex taste.

In the case of sound recognition the pattern theory may be described with reference to the operation of the cochlea. The dynamic physical parameters of the cochlea are such that the energy associated with different frequencies tends to localize at different points along the cochlea, high frequencies near the stapes and low frequencies near the helicotrema. The sensory structures convert this energy into a spatial pattern of loudness along the cochlea. The fundamental concept of the pattern theory of sound recognition is that sounds are recognized by the shape of this spatial pattern of loudness.

The recognition of a sustained sound is thus the recognition of a stationary pattern. The recognition of time varying sounds such as speech corresponds to the recognition of a three dimensional surface of
loudness in space and time, as shown in Figure 5.2. The phoneme, which is considered by many to be the basic unit of speech, is represented by a small section of this surface. The theory further hypothesizes that translation and distortion of patterns within limits does not greatly change the recognizability of the sound. This was demonstrated by Licklider in an investigation of speech word recognition as a function of various forms of distortion of the time and frequency axes (ref. 26). He found that the human has a limited range of adaptation in the recognition of speech.

In this dissertation the primary investigation concerns sustained patterns, which constitutes the first step in the solution to the more general problem presented by time varying patterns.

5.3 Concepts of Pattern Analysis and Recognition

In discussing the recognition of cochlear patterns there are several concepts which, if introduced, make the structure of the theory more coherent. A common artifice in communications theory is the construction of a geometrical signal model in abstract multi-dimensional space. Such a model will be introduced here to provide a clarification of the analysis and recognition processes.

Every waveform of finite length and finite frequency spectrum may be represented as a point or radius vector in multi-dimensional space. A signal, defined as a desired input of a communication system, is generally observed to lie within some specified time interval and is mixed with
FIGURE 5.2 Three Dimensional Surface for Recognition of Speech
undesired inputs termed noise. The noise is considered to enter the system independently of the signal and to affect observations according to how it and the signal are combined. Distortion of the signal may exist at the source due to errors in signal generation. This distortion may be considered as a noise which enters the system at the signal source.

Using a geometrical model, the analysis of a signal may be thought of in terms of a mapping process, in which the signal plus the noise is mapped onto a multi-dimensional signal space. In this space the signal plus noise is represented by a region about the point that represents the signal alone. The size of this region reflects the uncertainty about the signal due to the noise.

It is obvious that there are many permissible mappings of a particular ensemble of waveforms. The choice of a signal space and a particular mapping are determined by the characteristics of the signal and the desired analysis process.

Assume that, in the case of cochlear patterns, the source must draw the signal from a limited ensemble of waveforms. Distortion of the signal at the generator is then defined as noise entering the system at the signal source. The initial analysis will assume that, except at the source, the environment is noiseless. This assumption is obviously valid only if the signal to noise ratio is relatively large. The addition of random noise into the analysis and recognition process will be briefly considered later. All possible distortions of the limited ensemble of signals form the ensemble
of all-possible signals from the source. In the case of speech this distortion
is the variation in sound formation by the speaker.

The cochlear signal-to-pattern transformation represents the mapping
of the signal waveform into a pattern which is represented geometrically
as a point in multi-dimensional space. The limited ensemble of patterns
may be thought to be represented by regions in n-space, so that any point
in a region may be termed a selection from the limited pattern ensemble.
Thus a pattern from the ensemble of all-possible patterns will fall within
a region which is designated by a pattern from the limited ensemble.

The process of recognition is to pick members of the limited ensemble
of patterns to represent members of the all-possible pattern ensemble.
Thus with respect to a suitable error criterion, the nearest member of the
limited ensemble is used to designate a member of the all-possible ensemble.
If all of the information contained in the signal is drawn from the limited
ensemble the loss in information due to such a process is attributable only
to the error in correct recognition of members of the limited ensemble.

Assume that the mapping of a known limited pattern ensemble onto a
signal space is known. Then the recognition of an unknown pattern is the
determination of the region in space in which the unknown pattern is mapped.
If the choice is forced (that is, if it must be made in a specified time period),
the ensemble of all-possible patterns must be contained within the signal
space. Generally, however, neither the mapping nor the space are known,
and the relative position of an unknown pattern must be measured by deter-
mining its differences from each member of the limited ensemble. The unknown pattern is then recognized as, and represented by, the pattern of the nearest known point. This is shown in two dimensions in Figure 5.3. The region associated with each point of the limited ensemble is determined by the measurement of "nearness."

The word "nearness" in the preceding paragraph introduces the concept of a measure of distance between points in n-space—this measure in fact corresponds to an error criterion. In this connection, many different measures for "nearness" may be hypothesized. It is important to note that the relative weighting of the information-bearing components of the signal is determined by the mapping of the analysis process and the error criterion of the recognition process. For example, mean square error, in one n-space, which corresponds to absolute distance squared between points, may correspond to a magnitude error criterion for some other n-space.

In designing communication and control systems there is a good deal of discussion of optimization procedures. But it quickly becomes apparent that "optimum" must be defined with respect to a particular error criterion by which the errors are weighted. The choice of an optimum mapping for a recognition process is dependent on the relative importance of the information-carrying components of the signal. If the entire signal is not information bearing then the mapping must be selected to emphasize the information-bearing parts and to de-emphasize the non-information-bearing
FIGURE 5.3
parts. For example, consider the case of speech. It is possible to map speech waveforms onto a number of different signal spaces: one might be the spectrum of speech sounds, another might be a parametric description of the vocal tract, and a third might be the cochlear pattern of loudness. These three mappings all apply to the same waveform, but for the same error criterion they do not necessarily weight the relative importance of the information bearing components of the signal in the same manner.

In the study of human sound recognition, the hypothesized analysis and recognition process must ultimately be compared with human performance under similar conditions. The choice of a signal space for a sound recognition system is a very difficult one—the exact analytical solution for an optimum mapping is probably impossible. There is, however, a system for which certain parallels can be drawn which has undergone optimization over a long period of time—the human hearing system. By constructing an analysis and recognition system similar to the human it should be possible to achieve some form of optimization for a speech recognition system.

5.4 Cross-correlation as a Recognition Process

5.4.1 Cauchy-Schwarz Inequality

In this section the Cauchy-Schwarz Inequality for real integrals is derived. It will be obvious that in this form the inequality is basic to the cross-correlation recognition process.
Consider two functions, \( f(x) \) and \( g(x) \), which are real, integrable, and integrable absolute square over the interval from zero to \( L \). Then form the integral

\[
\int_0^L \left[ f(x) + \lambda g(x) \right] \left[ f(x) + \lambda^* g(x) \right] \, dx \tag{5.4.1}
\]

where \( \lambda \) is a complex number and \( \lambda^* \) is its conjugate. The integrand of this integral is positive and real, as it is the product of a function and its conjugate. Thus the integral is real, positive, and satisfies the inequality

\[
\int_0^L \left[ f(x) + \lambda g(x) \right] \left[ f(x) + \lambda^* g(x) \right] \, dx \geq 0 \tag{5.4.2}
\]

which may be expanded as

\[
\int_0^L \left[ f(x) \right]^2 \, dx + \left[ \lambda + \lambda^* \right] \int_0^L f(x) g(x) \, dx + \lambda \lambda^* \int_0^L \left[ g(x) \right] dx \geq 0 \tag{5.4.3}
\]

If \( \lambda \) and \( \lambda^* \) are expressed as

\[
\lambda = \text{Re}\lambda + j \text{Im}\lambda \tag{5.4.4}
\]

\[
\lambda^* = \text{Re}\lambda - j \text{Im}\lambda \tag{5.4.5}
\]

the product may be formed as

\[
\lambda \lambda^* = \left[ \text{Re}\lambda \right]^2 + \left[ \text{Im}\lambda \right]^2 = |\lambda|^2 \tag{5.4.6}
\]
Using these expressions, the inequality of 5.4.3 becomes

\[ \int_0^L [f(x)]^2 + 2 \Re \lambda \int_0^L f(x) g(x) \, dx + [\Re \lambda] \int_0^L [g(x)]^2 \, dx \]

\[ + [\Im \lambda]^2 \int_0^L [g(x)]^2 \, dx \geq 0 \quad (5.4.7) \]

Now consider the integral function of

\[ \mathcal{F}(\Re \lambda) = \int_0^L \left( f(x) + \Re \lambda \, g(x) \right)^2 \, dx \quad (5.4.8) \]

which certainly satisfies the inequality

\[ \int_0^L \left( f(x) + \Re \lambda \, g(x) \right)^2 \, dx \geq 0 \quad (5.4.9) \]

which expands as

\[ \int_0^L [f(x)]^2 \, dx + 2 \Re \lambda \int_0^L f(x) g(x) \, dx \]

\[ + [\Re \lambda]^2 \int_0^L [g(x)]^2 \, dx \geq 0 \quad (5.4.10) \]

The inequality is made up of the first three terms of the inequality 5.4.7, which is thus an even stronger inequality. To investigate the roots of the function of 5.4.8, identify the integrals of 5.4.10 as
\[ J(\text{Re}\lambda) = c + b \text{Re}\lambda + a \left[ \text{Re}\lambda \right]^2 \geq 0 \]  

Consider the plot of \( J(\text{Re}\lambda) \) as a function of \( \text{Re} \lambda \) in Figure 5.4. The function \( J(\text{Re}\lambda) \) can not become negative, and therefore, it can at most become tangent to the abscissa, which occurs when \( J(\text{Re}\lambda) = 0 \). Thus the roots of \( J(\text{Re}\lambda) \) can not be real and distinct, but rather they must be real and equal or complex conugates. For this to be true the discriminant of the quadratic equation must be negative or zero as

\[ b^2 - 4ac \leq 0 \]  

By substitution this becomes the integral inequality

\[ 4 \left[ \int_0^L f(x)g(x) \, dx \right]^2 - 4 \int_0^L \left[ f(x) \right]^2 \, dx \int_0^L \left[ g(x) \right]^2 \, dx \leq 0 \]  

which directly gives the Cauchy-Schwarz Inequality for real integrals as

\[ \left[ \int_0^L f(x)g(x) \, dx \right]^2 \leq \int_0^L \left[ f(x) \right]^2 \, dx \int_0^L \left[ g(x) \right]^2 \, dx \]  

5.4.2 Cross-correlation function

Of interest here is the recognition of a particular cochlear pattern from an ensemble of possible patterns. For practical purposes in the preliminary consideration of recognition, the pattern is considered to be in a noiseless environment.
FIGURE 5.4
The integral within the brackets on the left side of the inequality of 5.4.14 may be identified as the cross-correlation function of \( f(x) \) and \( g(x) \) for zero delay between \( f(x) \) and \( g(x) \). This integral, the cross-correlation function, is

\[
\int_{0}^{L} f(x) g(x) \, dx
\]  

(5.4.15)

The normalized cross-correlation function is defined as

\[
\rho = \frac{\int_{0}^{L} f(x) g(x) \, dx}{\sqrt{\int_{0}^{L} \left[ f(x) \right]^2 \, dx \int_{0}^{L} \left[ g(x) \right]^2 \, dx}}
\]  

(5.4.16)

To investigate the application of these functions to cochlear pattern recognition, assume that \( f(x) \) is an unknown pattern which is to be recognized. Assume that \( g_m(x) \) is the \( m^{th} \) pattern from a dictionary of patterns, this dictionary containing \( M \) possible patterns. The dictionary then contains the ensemble from which all patterns are drawn, and to which all unknown patterns are to be compared for the process of identification that constitutes recognition. Thus

\[ f(x) \quad \text{the pattern to be recognized} \]

\[ g_m(x) \quad \text{the \( m^{th} \) pattern from a stored ensemble of \( M \) patterns} \]

Both functions are real, non-negative, and of spatial extent \( L \) as defined in section 5.1.
To investigate the cross-correlation function as a recognition function it is convenient to study the Cauchy-Schwarz Inequality for various conditions. First, let

\[ f(x) = k g_m(x) \quad (5.4.17) \]

where \( k \) is an arbitrary real constant. Substituting this into the Cauchy-Schwarz Inequality yields

\[
\left[ \int_{0}^{L} k g_m(x) g_m(x) \, dx \right] \leq \int_{0}^{L} \left[ k g_m(x) \right]^2 \, dx \int_{0}^{L} \left[ g_m(x) \right]^2 \, dx \quad (5.4.18)
\]

Simplifying this yields

\[
k^2 \left[ \int_{0}^{L} \left[ g_m(x) \right]^2 \, dx \right] \leq k^2 \left[ \int_{0}^{L} \left[ g_m(x) \right]^2 \, dx \right] \quad (5.4.19)
\]

Thus the left and right sides of the inequality are identical and the equal sign applies. The cross-correlation function is a maximum value when the two patterns are proportional, and the normalized cross-correlation function has its maximum value of 1. It must also be noted that the normalized cross-correlation function is independent of the magnitude of the two patterns being cross correlated. This may be demonstrated by computing \( \rho \) for two patterns of arbitrary magnitude, \( k_1 f(x) \) and \( k_2 g(x) \), as
which becomes

\[ \Phi(k_1 f, k_2 g_m) = \frac{\int_0^L k_1 f(x) k_2 g_m(x) \, dx}{\sqrt{\int_0^L [k_1 f(x)]^2 \, dx \int_0^L [k_2 g_m(x)]^2 \, dx}} \]  

(5.4.20)

Returning to the Cauchy-Schwarz Inequality of equation 5.4.14, it is necessary to define the auto-correlation of a pattern as the cross-correlation of a pattern with itself. This is also the integral of a square magnitude and thus proportional to energy. From this the right side of equation 5.4.14 may be thought of as the product of the energy of \( f(x) \) and the energy of \( g_m(x) \).

Since the function \( g_m(x) \) is to be stored in a dictionary it is possible to set certain amplitude conditions on it. Thus adjust all of the \( g_m(x) \) so that they have equal energy, as

\[ \int_0^L [g_m(x)]^2 \, dx = A^2 \]  

(5.4.22)

where \( A \) is a constant for all \( M \) stored patterns.
This effectively normalizes all of the patterns stored in the dictionary.

Consider the unknown pattern \( f(x) \). In any particular case of recognition this pattern will have total energy set by the incoming sound magnitude. Thus

\[
\int_0^L [f(x)]^2 \, dx = B^2
\]  

(5.4.23)

where \( B \) is a constant for any particular pattern to be recognized. Using the Cauchy-Schwarz Inequality and the conditions of equations 5.4.22 and 5.4.23 there is obtained a further inequality yielding recognition. This is

\[
\int_0^L f(x) g_m(x) \, dx \leq AB = \sqrt{\int_0^L [f(x)]^2 \, dx} \sqrt{\int_0^L [g_m(x)]^2 \, dx}
\]

(5.4.24)

Thus the cross-correlation function on the left will, if computed for all of the \( M \) patterns in the dictionary, be a maximum for the one which is proportional to the unknown pattern.

Now envision a device that computes the correlation function of \( f(x) \) and each of the \( g_m(x) \) and then selects that \( g_m(x) \) showing the maximum correlation function. This is recognition! Note that the only normalization required is that of the stored patterns, a comparatively easy task. It is obvious that this sort of recognition could be accomplished in either a parallel or series manner, the parallel system operating with duplication of processing equipment and the series system operating with sampling
and signal storage. The block diagram of such a parallel elementary recognition device is shown in Figure 5.5

5.4.3 Mean square error

To this point in the discussion of the cross-correlation function as a pattern recognition function only the case has been considered where \( f(x) \) corresponds to one of the \( g_m(x) \) stored patterns. Now consider the problem where \( f(x) \) does not exactly correspond to any of the \( g_m(x) \) due to distortion, noise, or an insufficient dictionary of patterns. Which pattern will be recognized?

Consider the mean square error between \( f(x) \) and \( g_m(x) \) in the form

\[
\varepsilon^2 = \int_0^L \left[ f(x) - g_m(x) \right]^2 \, dx \quad (5.4.25)
\]

Expand this and rewrite it as

\[
\varepsilon^2 = \int_0^L \left[ f(x) \right]^2 \, dx + \int_0^L \left[ g_m(x) \right]^2 \, dx - 2 \int_0^L f(x) \, g_m(x) \, dx \quad (5.4.26)
\]

Since \( \varepsilon^2 \) can only be positive or zero this yields the conditions

\[
\begin{align*}
\varepsilon^2 &> 0 \quad \text{If } f(x) \neq g_m(x) \\
\varepsilon^2 &= 0 \quad \text{If } f(x) = g_m(x)
\end{align*}
\]

Again using the pattern energies given by equations 5.5.14 and 5.5.15 and the condition that all \( g_m(x) \) have the same energies, this equation becomes
Elementary Recognition Device

Device to Pick Peak Input for Recognition

Recognized Pattern

Cross-correlation

Peaks

Peaks

Indication of
\[ \epsilon^2 = B^2 + A^2 - 2 \int_0^L f(x) g_m(x) \, dx \]  
(5.4.27)

An inspection of this equation yields several interesting conclusions. First the mean square error and the correlation are functionally related, and thus by computing one the other may be determined. When one is a minimum the other is a maximum and vice versa. So instead of picking the maximum correlation function the \( g_m(x) \) with minimum mean square error could just as well be picked.

The main question of this sub-section may then be answered. If \( f(x) \) is not identical to one of the \( g_m(x) \), the elementary recognition device will select that \( g_m(x) \) which shows the least mean square error with \( f(x) \).

Therefore, in addition to the elementary recognition device of Figure 5.5 an equivalent device may be hypothesized which measures the mean square error between the incoming pattern and each of the stored patterns and picks as the recognized pattern the one having the least mean square error.

Another question needs to be answered. What is the minimum mean square error to be expected if the patterns are of the same shape but of different amplitude. To investigate this, return to equation 5.5.18 and assume that \( f(x) \) and \( g_m(x) \) differ only by a constant of proportionality, as

\[ f(x) = Kg_m(x) \]  
(5.4.28)
Then equation 5.5.18 becomes

\[ \epsilon^2 = \int_0^L k^2 \left[ g_m(x) \right]^2 dx + \int_0^L g_m(x)^2 dx \]

\[ - 2 \int_0^L k g_m(x)^2 dx \]  \hspace{1cm} (5.4.29)

Identifying the integral as the energy of \( g_m(x) \) this becomes

\[ \epsilon^2 = k^2 A^2 + A^2 - 2kA^2 = A^2 (K^2 + 1 - 2k) \]  \hspace{1cm} (5.4.30)

\[ = A^2 (K-1)^2 \]

This is the minimum mean square error to be expected for a particular pattern under the above conditions. As expected, if \( k = 1 \), then \( \epsilon = 0 \).

5.5 Absolute Difference as a Recognition Process

In the preceding section it was shown that recognition could be achieved with the cross-correlation function, and that this function led to a mean square error criterion. It was also noted that the mechanism of the central nervous system can not yet be hypothesized realistically. Another possible recognition function may be suggested by the inhibitory action of neural junctions. This function may have some disadvantages when compared to the cross-correlation function, although, it does offer a certain simplicity in large scale parallel operating systems such as may be represented by the central nervous system.
In certain neural junctions, two signals are combined and the output is a function of the two inputs in a manner similar to digital computer logic circuits. In such a junction one signal is inhibited by the other. If it is hypothesized that many inhibitory circuits are suitably interconnected, it may be possible to obtain an absolute difference between two patterns. A mechanism for a recognition process may thus be suggested where the unknown pattern and a stored pattern are compared on an absolute difference basis. The hypothesized recognition function requires minimization of

\[ \int_0^L \left| f(x) - g_m(x) \right| \, dx \]  

The problem associated with this recognition process is that finding a normalizing function has not yet been achieved—recognition must certainly not be excessively amplitude dependent. There are, however, two plausible functions, the loudness of the unknown pattern and the loudness of the stored pattern. The former may in fact be available to the central nervous system because it is subjective loudness. It can be hypothesized that the latter might be taken from storage, be amplitude adjusted to a value near that of the unknown pattern, and then used for recognition. This reasoning leads to suggesting the sum of the two loudness values (or possibly some function of the sum) as a normalizing function. A recognition process may thus be defined as
Little else can be done with this function on a mathematical basis, but it may readily be studied experimentally. This will be discussed in the next chapter.

The foregoing constitutes only one of the many functions that can be hypothesized for the recognition process. Much further study of human behavior must be made before relations can be hypothesized or approximated in a realistic manner.

5.6 Recognition of Cochlear Patterns

In this chapter the analysis and recognition of cochlear patterns has been discussed with reference both to recognition devices and to the human hearing system. And while here the interest is in hypothesizing a process that may be related to the human, it is obvious that in doing so the fundamental basis is discussed for a recognition device which might be incorporated into a general speech communication system. Thus the two devices, human and machine, may to a degree be treated interchangeably throughout this discussion.
Cross-correlation as a central nervous system processing technique has been suggested by Cherry (ref. 17), Reichardt (ref. 27) and others. It is generally admitted by each that, other than response, there are no known specific central nervous circuits that specifically indicate existence of such processes. However, this is not surprising in view of the very limited knowledge of the central nervous system.

The absolute difference recognition process was introduced from thoughts that neuron junctions may, by inhibitory combination, form differences of incoming and stored patterns. From such differences it is possible to hypothesize a simple recognition process. However, there comes at this point the problem of normalization so that recognition will be relatively amplitude independent. This problem has bothered the speech communication system designers a great deal. The principal advantage of the cross-correlation recognition process is that normalization may be conveniently accomplished in the pattern storage dictionary.

In hypothesizing analysis and recognition processes, the error criteria are not actually known and thus may not be used to establish the correct relations. As more data becomes available, it may be possible to compare the human and machine recognition processes so as to evaluate human error criteria.
CHAPTER 6
COCHLEAR PATTERN ANALYSIS OF EIGHT VOWELS

6.1 Introduction

In Chapter 5 the theoretical concepts of cochlear pattern recognition were developed to the extent necessary for sustained sounds. In this chapter these techniques will be applied to eight sustained vowels and the results compared with those for human recognition.

Voiced vowels are produced in the speaking apparatus of the human by shaping, with three cavities, a sound vibrational spectrum produced by the vocal folds (cords). It has been shown that the vowels may be described in terms of three bands of energy. Each band is termed a formant. The formants are called first, second, and third, in the order of their center frequencies. Each formant is associated with one of the sound forming cavities. For the sustained vowels the third, or highest frequency formant, has the lowest energy content of the three. It is also more likely to be masked than the second, or middle frequency formant.

The vocal folds vibrate at a fundamental frequency with an approximately triangular waveform. This produces a wide spectrum for shaping into sounds. Measurements indicate that this spectrum is not a line spectrum, but is broadened by randomness in the vocal fold excitation. Whispered
vowels are produced by relaxing the vocal folds, and using instead a rush of air through the pharynx as sound excitation for the shaping cavities. Although the whisper excitation spectrum is similar to white noise, the spectral shape of whispered vowels is not greatly different from that of voiced vowels. As will be shown later, the recognition of voiced and whispered vowels is not noticeably different.

6.2 Eight Selected Sustained Vowels

6.2.1 The selected vowels

Eight vowels were selected to be used in experimental studies of recognition. The selection was based on the desire to have a limited sample size which was representative of the possible range of sustained speech sounds. Those selected are given below in both phonetic symbol form and in a form which more literally indicates their sound. To improve clarity, the parenthetical form will be attached to the phonetic symbol throughout this dissertation.

<table>
<thead>
<tr>
<th>PHONETIC SYMBOL</th>
<th>APPROXIMATE SOUND</th>
<th>EXAMPLE OF SOUND</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>(ee) as in eat</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>(i) as in it</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>(eh) as in bet</td>
<td></td>
</tr>
<tr>
<td>æ</td>
<td>(a') as in can</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>(ah) as in father</td>
<td></td>
</tr>
<tr>
<td>/</td>
<td>(uh) as in but</td>
<td></td>
</tr>
</tbody>
</table>
The formant structure may be used to graphically display the vowels by plotting normalized values of the first and second formants. The frequency of the third formant is used as the normalizing factor. It should be noted that the amplitudes of the formants are not shown on this plot, and thus the vowel is not completely defined by the plot. Figure 6.1, after Peterson (ref. 28), presents such a plot for the eight sustained vowels selected for use in the present recognition studies. It is important to note that the plot does not differ greatly for men, women, and children. This is consistent with human ability to recognize speech sounds with little regard to the speaker. From the previous discussion of the cochlea, Figure 3.5 indicates that frequency localization along the cochlea is logarithmic. The relative formant similarity between like sounds by different speakers indicates that the shape of the patterns from the cochlea should be quite similar. However, the absolute position of the pattern will be shifted depending on the frequencies of the formants. Studies performed on the analog ear indicate that the shift in pattern position is small.

6.2.2 Human confusion matrices

A program was initiated by Brooks and Dimmick (ref. 29) to experimentally study the process of sound recognition as performed by the human. The eight selected vowels were recorded, both voiced and whispered. A tape was then made to present, in random order, 48 voiced (whispered)
A Plot of the Ratio of the First Three Formants for Approximately Phonetically Equivalent Sustained Vowels Pronounced by a Man, a Woman, and a Child (after G. E. Peterson, ref. 28)

FIGURE 6.1
vowels, each vowel being presented six times. The observers were presented with samples of about one second duration, and were instructed to make a check mark indicating which vowel was recognized.

This tape was presented to a total of 32 trained observers and the results tabulated into confusion matrices, one for voiced vowels and one for whispered vowels. These confusion matrices are shown in Figures 6.2 and 6.3.

6.3 **Vowel Patterns From the Analog Ear**

6.3.1 **Introduction**

A tape was made using the same speaker as used for the human recognition studies. This tape had longer samples, of about seven seconds each, so as to implement oscilloscope observations. The tape was fed into the analog ear and the patterns were photographed with a Polaroid oscilloscope camera. Camera shutter speed was maintained at 1/100 second, and the lens was set at f2.8. In some of the studies on the analog ear the lens opening was varied to emphasize or de-emphasize certain patterns in multiple exposure photographs.

A true RMS meter was used to monitor the patterns so that, within the operator's ability to trigger the camera release at the proper time, the patterns had approximately the same energy. Conditions were maintained constant throughout a series of photographs, and at least three samples of each vowel, voiced and whispered, were taken. The photographs were taken
VOICED VOWEL RECOGNITION TEST

CONFUSION MATRIX

<table>
<thead>
<tr>
<th>VOICED VOWEL - RESPONSE</th>
<th>i (ee)</th>
<th>I (1)</th>
<th>E (eh)</th>
<th>a (ah)</th>
<th>A (uh)</th>
<th>O (aw)</th>
<th>U (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>190</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>4</td>
<td>154</td>
<td>20</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>7</td>
<td>162</td>
<td>19</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td></td>
<td>4</td>
<td>173</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>1</td>
<td>146</td>
<td>10</td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
<td>1</td>
<td>7</td>
<td>28</td>
<td>153</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>U</td>
<td></td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>182</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1536 vowel presentations
32 observers - 8 vowels - each vowel presented 6 times
1342 correct recognitions
11 unidentified presentations

FIGURE 6.2
WHISPERED VOWEL RECOGNITION TEST

CONFUSION MATRIX

<table>
<thead>
<tr>
<th></th>
<th>i (ee)</th>
<th>I (i)</th>
<th>E (eh)</th>
<th>A (a)</th>
<th>A (ah)</th>
<th>A (uh)</th>
<th>O (aw)</th>
<th>U (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>188</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>159</td>
<td>32</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>11</td>
<td>131</td>
<td>47</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>176</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
<td>132</td>
<td>18</td>
<td>35</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>18</td>
<td>169</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
<td>7</td>
<td>1</td>
<td>180</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>191</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1536 vowel presentations
32 observers - 8 vowels - each vowel presented 6 times
1326 correct recognitions
5 unidentified presentations

FIGURE 6.3
without the smoothing circuits of the sampling system so as to indicate the actual sample points along the cochlea.

6.3.2 Quantization and sampling

For each vowel, the pattern was manually traced and smoothed from three photographic patterns and the average of these used as the pattern representing that vowel. Although there was little difference between samples, this averaging was carried out to insure that a representative sample was obtained.

To perform the analysis required for recognition without the aid of continuous devices it was necessary to use samples from the patterns, (where the number of these samples bears no relationship to the 36 samples taken by the commutator). The patterns were also quantized and the number of levels was selected with due consideration of the accuracy to which the pattern could be measured. A magnitude scale of 20 units was selected with quantization to 1/2 unit to yield 40 magnitude levels quantization. It will be noted that the magnitude scaling is arbitrary and chosen, in part, for convenience in manipulation in the recognition processing.

To decide how many samples from a pattern were necessary to adequately describe the pattern, the energy for several patterns was computed for various numbers of samples. It was noted that, as the sample size was increased beyond a certain point, there was little change in the pattern energy. Thus a sampling was selected such that the energy was relatively independent of the sampling.
The procedure selected was to sample the patterns at 19 points along the cochlea—the end points, which are necessarily zero, are included in the tabulations for clarity. From the theory of sampled data the following estimates are used for the continuous pattern parameters.

Let \( a_i, b_i, \text{ etc.} \), be the \( i \)th pattern sample, where, in this case, \( i = 0, 1, 2, \ldots 21 \). Then

\[
\begin{align*}
a_0 &= a_{21} = 0 \\
a_i &\neq 0
\end{align*}
\]

**Pattern Area**

\[
\sum_{n=0}^{21} a_i \quad (6.3.1)
\]

**Pattern Energy**

\[
\sum_{n=0}^{21} a_i^2 \quad (6.3.2)
\]

**Pattern Cross-correlation Function**

(for patterns with samples \( a_i \) and \( b_i \) respectively)

\[
\sum_{n=0}^{21} a_i b_i \quad (6.3.3)
\]
**Pattern Normalized Cross-correlation Coefficient**

\[
P = \frac{\sum_{n=0}^{21} a_i b_i}{\sqrt{\sum_{n=1}^{21} a_i^2 \sum_{n=1}^{21} b_i^2}}
\]

(6.3.4)

**Pattern Absolute Difference**

\[
\sum_{n=1}^{21} |a_i - b_i|
\]

(6.3.5)

**Pattern Normalized Absolute Difference**

\[
\frac{\sum_{n=1}^{21} |a_i - b_i|}{\sum_{n=1}^{21} a_i + \sum_{n=1}^{21} b_i}
\]

(6.3.6)

6.3.3. **Vowel pattern data**

The cochlear patterns of vowels used in the recognition studies are shown in Figures 6.4 through 6.11. The actual data from the average patterns is indicated by heavy dots, while the quantized pattern is outlined. The
E(6A) VOICED
E = 958

P = 0.977

E(6A) WHISPERED
E = 1035

FIGURE 6.6  DISTANCE
pattern energy is indicated in the upper right-hand corner of each pattern. The normalized cross-correlation coefficient between the voiced and whispered pattern is also given on the set of patterns for each sound. Vowel pattern data are also tabulated in Tables 6.1 and 6.2.

6.4 Cross-correlation of Eight Selected Vowels

6.4.1 Correlation matrices

The normalized cross-correlation coefficient was computed for the voiced and whispered patterns for each vowel. The results of these computations are shown in the correlation matrix of Figure 6.12. This matrix shows a normalized cross-correlation coefficient of about 0.9 or greater in all cases except one. This exception is quite interesting and will be discussed later.

The normalized cross-correlation coefficient was computed for all pairs of voiced and of whispered vowels. From these computations, the correlation matrices were constructed. The matrix is shown for voiced vowel patterns in Figure 6.13, and for whispered vowel patterns in Figure 6.14. Note that these matrices are symmetrical about the diagonal, therefore, only half of the matrix is shown.

It will be noted from the matrices that the nearer the vowels are together on the formant plot of Figure 6.1 the more nearly they correlate.

6.4.2 Recognition matrices and recognition scatter diagram

The information of the preceding sections is summarized in


<table>
<thead>
<tr>
<th>Sample Number</th>
<th>i (ee)</th>
<th>I (i)</th>
<th>E (eh)</th>
<th>a (a)</th>
<th>ah (ah)</th>
<th>u (uh)</th>
<th>o (aw)</th>
<th>oo (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
<td>4.0</td>
<td>3.5</td>
<td>2.0</td>
<td>4.0</td>
<td>2.0</td>
<td>4.0</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>5.0</td>
<td>7.0</td>
<td>5.0</td>
<td>4.0</td>
<td>4.0</td>
<td>7.0</td>
<td>3.0</td>
<td>6.0</td>
</tr>
<tr>
<td>5</td>
<td>8.0</td>
<td>10.5</td>
<td>9.0</td>
<td>6.5</td>
<td>5.0</td>
<td>11.0</td>
<td>6.0</td>
<td>10.0</td>
</tr>
<tr>
<td>6</td>
<td>15.0</td>
<td>14.0</td>
<td>13.0</td>
<td>10.0</td>
<td>7.0</td>
<td>14.0</td>
<td>8.5</td>
<td>14.0</td>
</tr>
<tr>
<td>7</td>
<td>16.0</td>
<td>14.5</td>
<td>14.0</td>
<td>12.5</td>
<td>7.5</td>
<td>12.0</td>
<td>8.0</td>
<td>12.0</td>
</tr>
<tr>
<td>8</td>
<td>13.0</td>
<td>13.5</td>
<td>14.0</td>
<td>14.5</td>
<td>8.0</td>
<td>10.0</td>
<td>6.0</td>
<td>10.0</td>
</tr>
<tr>
<td>9</td>
<td>5.0</td>
<td>9.5</td>
<td>11.5</td>
<td>15.0</td>
<td>10.0</td>
<td>9.0</td>
<td>5.5</td>
<td>9.5</td>
</tr>
<tr>
<td>10</td>
<td>2.0</td>
<td>7.0</td>
<td>8.5</td>
<td>14.0</td>
<td>15.0</td>
<td>11.5</td>
<td>8.5</td>
<td>10.5</td>
</tr>
<tr>
<td>11</td>
<td>0.5</td>
<td>4.5</td>
<td>5.5</td>
<td>10.0</td>
<td>17.0</td>
<td>10.5</td>
<td>12.0</td>
<td>11.5</td>
</tr>
<tr>
<td>12</td>
<td>0.0</td>
<td>3.0</td>
<td>4.0</td>
<td>8.0</td>
<td>15.5</td>
<td>7.5</td>
<td>14.5</td>
<td>10.0</td>
</tr>
<tr>
<td>13</td>
<td>0.5</td>
<td>2.5</td>
<td>3.5</td>
<td>7.0</td>
<td>7.0</td>
<td>5.0</td>
<td>14.0</td>
<td>6.0</td>
</tr>
<tr>
<td>14</td>
<td>0.5</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>6.0</td>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.0</td>
<td>3.0</td>
<td>1.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.5</td>
<td>3.0</td>
<td>4.0</td>
</tr>
<tr>
<td>16</td>
<td>1.0</td>
<td>2.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>3.0</td>
</tr>
<tr>
<td>17</td>
<td>0.5</td>
<td>1.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>2.0</td>
</tr>
<tr>
<td>18</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>19</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.5</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>21</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Pattern Energy

Energy  779  963  958  1177  1130  1061  922  1112

Pattern Area

Area  72.5  101.0  99.5  112.5  106.5  115.5  101.5  120.5

Average energy 1013

TABLE 6.1
### WHISPERED VOWELS

#### Quantized Pattern Shape Data

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>(ee)</th>
<th>(eh)</th>
<th>(a)</th>
<th>(ah)</th>
<th>(uh)</th>
<th>(aw)</th>
<th>(oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>2.0</td>
<td>2.5</td>
<td>2.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>3.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>7.0</td>
<td>6.0</td>
<td>7.0</td>
<td>5.5</td>
<td>2.0</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>12.5</td>
<td>10.0</td>
<td>10.0</td>
<td>8.0</td>
<td>3.0</td>
<td>2.5</td>
<td>2.0</td>
</tr>
<tr>
<td>6</td>
<td>18.0</td>
<td>16.5</td>
<td>15.0</td>
<td>9.0</td>
<td>4.5</td>
<td>4.0</td>
<td>3.5</td>
</tr>
<tr>
<td>7</td>
<td>15.0</td>
<td>16.0</td>
<td>16.0</td>
<td>10.0</td>
<td>6.0</td>
<td>6.0</td>
<td>5.5</td>
</tr>
<tr>
<td>8</td>
<td>6.0</td>
<td>8.0</td>
<td>15.0</td>
<td>10.5</td>
<td>8.5</td>
<td>9.0</td>
<td>7.0</td>
</tr>
<tr>
<td>9</td>
<td>1.5</td>
<td>3.0</td>
<td>9.0</td>
<td>10.5</td>
<td>12.0</td>
<td>13.0</td>
<td>10.0</td>
</tr>
<tr>
<td>10</td>
<td>0.5</td>
<td>1.0</td>
<td>5.0</td>
<td>10.5</td>
<td>15.0</td>
<td>18.0</td>
<td>14.0</td>
</tr>
<tr>
<td>11</td>
<td>0.0</td>
<td>0.5</td>
<td>4.5</td>
<td>11.5</td>
<td>15.0</td>
<td>17.0</td>
<td>16.0</td>
</tr>
<tr>
<td>12</td>
<td>0.0</td>
<td>0.5</td>
<td>3.5</td>
<td>11.0</td>
<td>7.0</td>
<td>13.0</td>
<td>15.0</td>
</tr>
<tr>
<td>13</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>3.0</td>
<td>1.5</td>
<td>7.0</td>
<td>10.0</td>
</tr>
<tr>
<td>14</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>1.5</td>
<td>3.0</td>
</tr>
<tr>
<td>15</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>16</td>
<td>1.5</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>17</td>
<td>2.0</td>
<td>1.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>18</td>
<td>2.5</td>
<td>2.0</td>
<td>1.5</td>
<td>1.5</td>
<td>1.0</td>
<td>2.0</td>
<td>0.5</td>
</tr>
<tr>
<td>19</td>
<td>3.0</td>
<td>3.0</td>
<td>2.0</td>
<td>1.5</td>
<td>1.0</td>
<td>2.5</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>3.0</td>
<td>3.0</td>
<td>2.0</td>
<td>1.5</td>
<td>1.0</td>
<td>2.5</td>
<td>1.0</td>
</tr>
<tr>
<td>21</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

#### Pattern Energy

<table>
<thead>
<tr>
<th>Sample</th>
<th>Energy (900)</th>
<th>786</th>
<th>1035</th>
<th>891.7</th>
<th>794</th>
<th>1163.9</th>
<th>986</th>
<th>1095</th>
</tr>
</thead>
</table>

#### Average Energy 956

**TABLE 6.2**
## CROSS-CORRELATION MATRIX OF VOICED AND WHISPERED VOWELS

FOR THE NORMALIZED CROSS-CORRELATION COEFFICIENT

<table>
<thead>
<tr>
<th></th>
<th>( i' ) (ee)</th>
<th>( I ) (i)</th>
<th>( E ) (eh)</th>
<th>( Æ ) (a)</th>
<th>( æ ) (ah)</th>
<th>( A ) (uh)</th>
<th>( Ω ) (aw)</th>
<th>( u ) (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i' ) (ee)</td>
<td>0.895</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( I ) (i)</td>
<td>0.901</td>
<td>0.977</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( E ) (eh)</td>
<td>0.962</td>
<td>0.944</td>
<td>0.827</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Æ ) (a)</td>
<td></td>
<td>0.977</td>
<td>0.944</td>
<td>0.922</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( æ ) (ah)</td>
<td></td>
<td>0.962</td>
<td>0.944</td>
<td>0.922</td>
<td>0.927</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( A ) (uh)</td>
<td></td>
<td>0.977</td>
<td>0.944</td>
<td>0.922</td>
<td>0.927</td>
<td>0.927</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Ω ) (aw)</td>
<td></td>
<td>0.962</td>
<td>0.944</td>
<td>0.922</td>
<td>0.927</td>
<td>0.927</td>
<td>0.927</td>
<td></td>
</tr>
<tr>
<td>( u ) (oo)</td>
<td></td>
<td>0.901</td>
<td>0.977</td>
<td>0.944</td>
<td>0.922</td>
<td>0.927</td>
<td>0.927</td>
<td>0.927</td>
</tr>
</tbody>
</table>

**FIGURE 6.12**
CROSS-CORRELATION MATRIX - VOICED VOWEL PATTERNS
FOR THE NORMALIZED CROSS-CORRELATION COEFFICIENT

<table>
<thead>
<tr>
<th></th>
<th>ɪ (ee)</th>
<th>ɪ (i)</th>
<th>ɛ (eh)</th>
<th>ɛ (a)</th>
<th>a (ah)</th>
<th>a (uh)</th>
<th>o (aw)</th>
<th>u (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ɪ</td>
<td>1.00</td>
<td>0.95</td>
<td>0.92</td>
<td>0.77</td>
<td>0.52</td>
<td>0.82</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>ɪ</td>
<td>1.00</td>
<td>0.99</td>
<td>0.91</td>
<td>0.72</td>
<td>0.94</td>
<td>0.66</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>ɛ</td>
<td>1.00</td>
<td>0.95</td>
<td>0.78</td>
<td>0.95</td>
<td>0.76</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ɛ</td>
<td>1.00</td>
<td>0.91</td>
<td>0.94</td>
<td>0.78</td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>1.00</td>
<td>0.89</td>
<td>0.93</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>1.00</td>
<td>0.87</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o</td>
<td>1.00</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 6.13**
CROSS-CORRELATION MATRIX - WHISPERED VOWEL PATTERNS
FOR THE NORMALIZED CROSS-CORRELATION COEFFICIENT

<table>
<thead>
<tr>
<th>Whispered Vowels</th>
<th>i (ee)</th>
<th>I (i)</th>
<th>E (eh)</th>
<th>A (ah)</th>
<th>A (uh)</th>
<th>C (aw)</th>
<th>U (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i (ee)</td>
<td>1.00</td>
<td>0.96</td>
<td>0.86</td>
<td>0.64</td>
<td>0.46</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td>I (i)</td>
<td>1.00</td>
<td>0.87</td>
<td>0.71</td>
<td>0.46</td>
<td>0.39</td>
<td>0.35</td>
<td>0.87</td>
</tr>
<tr>
<td>E (eh)</td>
<td>1.00</td>
<td>0.88</td>
<td>0.69</td>
<td>0.63</td>
<td>0.49</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>A (ah)</td>
<td>1.00</td>
<td>0.92</td>
<td>0.90</td>
<td>0.88</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (uh)</td>
<td>1.00</td>
<td>0.96</td>
<td>0.93</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C (aw)</td>
<td>1.00</td>
<td>0.98</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U (oo)</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 6.14
this section by defining a recognition matrix. The recognition matrices for voiced and whispered vowels are shown in Figures 6.15 and 6.16. The recognition matrix is a combination confusion matrix, from Figure 6.2 or 6.3, and correlation matrix, from Figure 6.13 or 6.14. There is, however, one important difference in that vowel confusions that are symmetrical about the diagonal are combined by summation so that the recognition matrix has the form of a symmetrical matrix. This alters slightly the form of the recognition data. However, confusions which are symmetrical about the diagonal have identical normalized cross-correlation coefficients so it does not matter how two vowels are confused—the associated correlation coefficient remains unchanged.

The number in the top of each partition of the recognition matrix is the normalized cross-correlation coefficient. The number in the bottom of each partition is the number of recognitions from the human recognition tests.

To compare the results of human and machine recognition, the pairs of numbers from recognition matrices were plotted on the scatter diagram of Figure 6.17. An ideal recognition system correctly recognizes a pattern only when the normalized correlation coefficient is equal to one. The addition of noise into an ideal system has the effect of introducing errors. These errors are evidenced by the confusion of patterns having high correlation coefficients. Thus the theoretically correct curve should be
RECOGNITION MATRIX OF VOICED VOWELS USING CROSS-CORRELATION ANALYSIS

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\imath} ) (ee)</th>
<th>( \hat{\imath} ) (i)</th>
<th>( \hat{\varepsilon} ) (eh)</th>
<th>( \hat{\alpha} ) (a)</th>
<th>( \hat{\lambda} ) (ah)</th>
<th>( \hat{o} ) (aw)</th>
<th>( \hat{\alpha} ) (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\imath} )</td>
<td>1.00</td>
<td>0.95</td>
<td>0.92</td>
<td>0.77</td>
<td>0.52</td>
<td>0.82</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>190</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \hat{\imath} )</td>
<td>1.00</td>
<td>0.99</td>
<td>0.91</td>
<td>0.72</td>
<td>0.94</td>
<td>0.66</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>154</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>( \hat{\varepsilon} )</td>
<td>1.00</td>
<td>0.95</td>
<td>0.78</td>
<td>0.95</td>
<td>0.76</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>162</td>
<td>23</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>( \hat{\alpha} )</td>
<td>1.00</td>
<td>0.91</td>
<td>0.94</td>
<td>0.78</td>
<td>0.94</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>173</td>
<td>12</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \hat{\lambda} )</td>
<td>1.00</td>
<td>0.89</td>
<td>0.93</td>
<td>0.91</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>146</td>
<td>38</td>
<td>35</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( \hat{o} )</td>
<td>1.00</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>182</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \hat{\alpha} )</td>
<td>1.00</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>182</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
RECOGNITION MATRIX OF WHISPERED VOWELS

USING CROSS-CORRELATION ANALYSIS

<table>
<thead>
<tr>
<th></th>
<th>i (ee)</th>
<th>I (i)</th>
<th>E (eh)</th>
<th>e (a)</th>
<th>a (ah)</th>
<th>A (uh)</th>
<th>J (aw)</th>
<th>u (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>1.00</td>
<td>0.96</td>
<td>0.86</td>
<td>0.64</td>
<td>0.46</td>
<td>0.31</td>
<td>0.28</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>188</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>1.00</td>
<td>0.94</td>
<td>0.71</td>
<td>0.46</td>
<td>0.39</td>
<td>0.35</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>43</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>1.00</td>
<td>0.88</td>
<td>0.69</td>
<td>0.63</td>
<td>0.49</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>131</td>
<td>53</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>1.00</td>
<td>0.92</td>
<td>0.90</td>
<td>0.88</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>176</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>1.00</td>
<td>0.96</td>
<td>0.93</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>33</td>
<td>42</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1.00</td>
<td>0.98</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>169</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1.00</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>180</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>191</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 6.16
expected to have a shape such as that defined by the points on the scatter diagram.

6.5 Absolute Difference Analysis of Eight Selected Vowels

The recognition process discussed in section 5.5 was experimentally studied using data from the voiced vowel patterns. The normalized absolute difference was computed as

\[ \sum_{n=1}^{21} |a_i - b_i| \]

\[ \frac{\sum_{n=1}^{21} a_i + \sum_{n=1}^{21} b_i}{21} \]  

(6.5.1)

Using these values and the recognition data from the human confusion tests, the recognition matrix of Figure 6.18 was constructed. The paired numbers of this matrix were plotted on the scatter diagram of Figure 6.19.

This scatter diagram is not so simply interpreted as the previous one for cross-correlation because the recognition function is not amplitude independent. However, the distribution of points certainly suggests that, the lower the normalized absolute difference, the more likely the confusion between sounds. With suitable normalization for amplitude differences, this process nevertheless yields recognition.
**Recognition Matrix of Voiced Vowels**

**Using Absolute Difference Analysis**

<table>
<thead>
<tr>
<th></th>
<th>( i ) (ee)</th>
<th>( I ) (1)</th>
<th>( E ) (eh)</th>
<th>( \alpha ) (a)</th>
<th>( u ) (uh)</th>
<th>( O ) (aw)</th>
<th>( u ) (oo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>0.0</td>
<td>0.193</td>
<td>0.215</td>
<td>0.346</td>
<td>0.483</td>
<td>0.314</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>190</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( I )</td>
<td>0.0</td>
<td>0.092</td>
<td>0.232</td>
<td>0.351</td>
<td>0.150</td>
<td>0.358</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td>154</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>( E )</td>
<td>0.0</td>
<td>0.165</td>
<td>0.306</td>
<td>0.151</td>
<td>0.328</td>
<td>0.168</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>162</td>
<td>23</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.0</td>
<td>0.169</td>
<td>0.158</td>
<td>0.243</td>
<td>0.177</td>
<td>0.358</td>
<td>0.220</td>
</tr>
<tr>
<td></td>
<td>173</td>
<td>12</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( a )</td>
<td>0.0</td>
<td>0.230</td>
<td>0.173</td>
<td>0.272</td>
<td>0.085</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>146</td>
<td>38</td>
<td>35</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>182</td>
</tr>
<tr>
<td>( u )</td>
<td>0.0</td>
<td>0.153</td>
<td>0.182</td>
<td>0.0</td>
<td>0.221</td>
<td>0.221</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>182</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6.18*
SCATTER DIAGRAM

ABSOLUTE DIFFERENCE RECOGNITION

FIGURE 6.10

Normalized Absolute Difference
6.6 Vowel Pattern Similarity Among Speakers

A brief experimental study was made of the vowel patterns of several speakers. This was done to verify the accuracy of the patterns used in the analysis portion of this dissertation and to gain some idea of the variability of vowel patterns that might be expected. In Figures 6.20 and 6.21, the patterns from speaker No. 1 are those used in the recognition studies. The others, pattern sets No. 2 through No. 5, were made by trained speakers from the University of Arizona Speech Department. In the latter cases the vowel sound was made by one person and monitored by at least one other person. Multiple photographs were taken and the average of two of the patterns agreed upon by the maker and monitor were used.

It was interesting to note during this process of pattern collecting that the pattern from the analog ear seemed to serve as a more sensitive indication of the sound being produced than the judgment of an untrained listener. In other words, by watching the pattern and listening to the speaker it was possible to detect changes in the sound that were not normally noticed by an untrained listener.

The vowel patterns of Figures 6.20 and 6.21 in each column are generally quite similar, with three exceptions. For the vowel /uh/, speaker No. 1 shows a stronger second than first formant; this is in contrast to the formant structure of the other four speakers. Also the vowel /oo/ as made by the first and fifth speakers was observed to have
a stronger second formant than first. Subsequent tests indicated a lack of unanimous agreement as to how this vowel should sound.

With respect to this latter disagreement and the differences in patterns, it must be kept in mind that the mean square error criterion applied here weighs large errors more heavily than smaller ones, and thus small variations may occur without significant change in the correlation coefficient.

6.7 Discussion of Results

This chapter has been devoted to an experimental study of analysis and recognition processes. The scatter diagrams are thought to sufficiently define the results of these processes; in the two cases considered (cross-correlation and magnitude of difference) they indicate recognition. The problem of normalization makes the cross-correlation function the more easily applied of the two, although it is not necessarily the more nearly correct one from a neural point of view.

The study of vowel similarity among speakers furnished an interesting observation concerning one of the patterns used in the recognition analysis. It was noted in section 6.4.1 that the vowel \( \Lambda(\text{uh}) \) is not so highly correlated between the voiced and whispered cases as are the other vowels. In Figure 6.15 the vowels \( \alpha(\text{ah}) \) and \( \Lambda(\text{uh}) \) have a normalized cross-correlation coefficient of 0.89 and a high confusion rate with 38 recognitions. The point on the scatter diagram representing this recognition pair is somewhat
out of line with the bulk of the other points and can be questioned. After first observing the dis-similarity of the vowel pattern with the others of column three of Figure 6.21, an investigation was made of the patterns for this vowel from the master tape. It was observed that this vowel was not the same throughout the tape, and thus a new set of photographs was taken. The cross-correlation coefficient between $a(\text{ah})$ and the new $\text{uh}$ pattern was then computed to be 0.94. The new point is also indicated on the scatter diagram of Figure 6.17 with an arrow from the first point to the replacement point. The effect of the modified vowel is also noted on Figure 6.19, and is indicated in relation to the original point with an arrow in the direction of the change. Further checking of the other vowel patterns failed to indicate any other significant vowel variations.

In studying the data of this chapter, it must be kept in mind that it is the error criterion used in the recognition analysis that determines the amount of difference between patterns. If the sound-to-pattern transformation or the error criterion are changed, the relative positions of the sound will be changed. In terms of the geometrical concept of analysis and recognition, the analysis and recognition processes are determined by the mapping and the measure of "nearness."

Before a set of patterns can be adopted as a "standard" it will be necessary to complete an extended experimental study of sounds and patterns in which the patterns and their variability will have to be determined. From
this information the number of patterns necessary to characterize a particular communication process can be established. Only at this point will it be possible to completely define a system for communication which uses patterns and is able to match the channel and the system to the ultimate receiver—the human brain.
CHAPTER 7
SUMMARY AND RECOMMENDATION

7.1 Introduction

In the preceding chapters of this dissertation, development of an analog of the ear has been described and studies of the patterns of loudness using this analog have been detailed. Pattern studies were based on the analysis necessary for recognition and were both theoretical and experimental. In this chapter certain aspects of this work will be discussed and related to other topics of interest.

One of the apparent things about the work done elsewhere in speech analysis, and speech recognition in particular, is the emphasis on the speech-making system and its ability to furnish parametric description of sounds. Of course, sounds must ultimately be processed by the human ear and central nervous system. There are several reasons for this historic or customary approach. First, the speech making apparatus is more easily studied and analyzed than the ear; and second, the construction of a human sound synthesis analog is somewhat simpler than the human sound analysis analog.

There are a number of topics that might be discussed in relation to the structure of speech itself. One topic of interest, and to which no immediate
solution is known, is the accuracy with which speech sounds must be made. It would seem that the device in the complete communication system that is of prime importance, so far as recognition is concerned, is the receiver and its characteristics.

Another emphasis encountered in the literature is that of describing speech sounds in terms of their spectra. This seems a logical step in view of the ease with which such measurements may be performed. However, there is also reason to be suspicious of this idea, at least in its pure form. Consider, for example, a case where a message is to be transmitted between two parties who are able to read the same language. The message is encoded for transmission such that a readable language for the sender and for the receiver exists only at the end points. If ability to read the end point language is assumed, there is no reason to assume the ability to read the "channel language"—the encoding may be a non-one-to-one process for which the inversion is unknown. Although laborious, it may be possible in principle for an observer to build a structure which will, in terms of "channel language" received, reduce the uncertainty of the response of the receiver to a particular message from the sender.

Except in the limit the inverse operation of the encoder may not be known, and it may be burdened with weighting the statistics of the messages in some unreasonable manner. Complete understanding may, in this case, never be possible in a reasonable amount of time.
In a like manner acoustical speech is a "channel language" if the speech system is interpreted in terms of the receiver. The ultimate receiver is the human brain. It is believed that more emphasis should be placed on the end language and less on the channel language. The concept of cochlear patterns is believed to be a powerful approach to the subject of speech analysis and recognition. The problem is not altogether simple, but the concept combines work from a number of fields into a common approach.

The function of the cochlea is that of a frequency analysis device which operates on the spectrum of received sound in a complicated manner. Consider the mapping of the speech signal ensemble in multi-dimensional signal space. From the confusing literature concerning recognition when it is discussed in terms of spectra, it seems logical to assume that, in this particular signal space, there is a good deal of overlapping of the points of similar classes of speech sounds.

Now assume that a process exists by which this signal space, the spectrum, may be mapped into another multi-dimensional signal space. In this second signal space, the points of the sound classes are more separated and form clusters about pertinent mean points. Then the distance between points in the class has clearly been reduced, and recognizability has been increased. From observation of the recognition ability of the human, the cochlear pattern of loudness is believed to constitute such a mapping.
Another topic of interest concerns a very common expression encountered in subjects dealing with neural activity and the human brain. This is the dogma that "every bit of information sent to the brain is used." Does this really make such good sense, or is this tied to a dream of a perfect human? Of fundamental importance is that a communication system is involved, and it must be realized that such systems, be they mechanical, electronic, or biological, must function in a similar manner. In short, the game played with nature must have a set of fixed rules. It is true that every bit of information may be used in arriving at the statistics of a non-deterministic process. However, the loss of one sample from a large group does not appreciably change the statistics of the system. It is very important in discussing information processing systems to understand the manner in which units of information add to form the overall statistical information.

7.2 Discussion of the Research

The results of this research are quite broad and rather difficult to interpret. It was originally proposed that an analog of the ear be developed and built. This was done. It has been shown that this analog is remarkable in certain respects. It is not, nor is it necessarily meant to be, sufficient in all respects. There are many things about the ear that are not modeled. Some of these, such as neural volleys, are important and bear on the general transfer sound-to-pattern characteristics. Others, such as fatigue, are
important in certain studies, but are not of great importance in the work presented here. Thus the model which has been proposed and has been built as an analog is a simple abstraction of those parameters of the ear which were considered to be important at the outset.

The process of modeling, however, requires a constant comparison of the model and the measured real world. It must be kept in mind in any research endeavor that knowledge of the real world is limited by the ability to make measurements. In the case of the ear, measurements are extremely difficult to make. The introduction of an analog which can be compared to a physiological specimen or used to demonstrate psychoacoustical phenomena can make future measurements much more simple. These measurements must, of course, be verified through appeal to experimental evidence.

The present analog is one of the first such complete models of the ear, and while it has shown important similarity to the human ear, it is not to be thought as complete or finished. It will require subsequent studies and modifications to more closely approach a complete simulation for the human ear.

The investigation of the analysis processes for recognition indeed implies human-like characteristics using two different criteria. However, the actual function used in the human has not yet been demonstrated. Rather, the human error criterion and recognition processes can not be further clarified without considerably more experimentation of the sort which
involves the human system under various signal and noise conditions.

7.3 Conclusions

From section 1.2, the problem of this dissertation was stated as: Study the analysis techniques necessary to characterize the process of sound recognition, and to demonstrate that distinguishable sounds have distinguishable cochlear patterns. This has been done.

The fundamental hypothesis, namely that sounds are recognized by virtue of the patterns of motion created in the sensory structure of the cochlea, has been demonstrated as acceptable using two recognition function criteria. The additional hypothesis, that distinguishable sounds have distinguishable cochlear patterns, has been demonstrated for the case of certain sustained sounds (by direct experiment).

7.4 A Communication System Using Recognition

In chapter 2 the process of recognition was discussed from the viewpoint of a device to perform the process. If such a device were built it would be able to indicate what sound was being received. If this indication were in the form of a code, this code could be used as the channel code for a communication system. Storage of patterns at the transmitter and the receiver ("memory") would allow the signal to be encoded at the sending end, transmitted, and then recoded into sound at the receiver. This in fact would be a communication system.
There are a number of things that would have to be known in order to build such a system. First, a set of patterns would have to be found that contained sufficient data for conveying the necessary information in speech between sender and receiver. Next, a coding would have to be devised to transmit the pattern information in an optimum manner.

The bandwidth reduction that such a system might achieve is dependent on the amount of information that is to be conveyed. If only the basic elements of speech are to be transmitted, the compression might be quite impressive. If qualities such as speaker recognition were desired, the price would have to be paid with bandwidth. There are, however, several attractive aspects in addition to bandwidth, per se. Such a system does lead to the possibility of convenient digital coding, for which considerable work in optimum coding has been done. Also, it offers some promise for a secrecy system, where the coding process from pattern-to-channel can be done so that the inverse process is very difficult to perform without a priori knowledge of the encoding-decoding scheme.

7.5 Recommendations for Further Research

There are several topics for which it is believed further research efforts are desirable. Some of these are necessarily long term, while others may show worthwhile short-term advancements.

The further application of the viewpoint discussed in section 1.1 to biological systems is believed to be particularly worthwhile. With respect
to the ear, there is suggested extension of the present work by refining models to be used as a basis for research. It also involves building these refinements into the analog so that it displays more psycho-acoustic phenomena, and is thus a better analog of the ear.

With regard to pattern recognition, there are several studies that are considered to be worthwhile. First, the variability of vowels as made by human speakers should be investigated. Second, the development of a simple speech synthesizer and the determination of "standard vowels" for test purposes would be extremely useful. This could be done in a statistical manner using a number of trained listeners to pick best-vowel sounds. Third, the long term program of determining how many patterns are necessary to characterize speech sound, employing various error criteria, is necessary before a recognition type communication system can be built.
REFERENCES


2. Helmholtz, H., Die lehre von den Tonempfindungen, Braunschweig, Germany, 1877.


