PERCEPTION OF SPECTRALLY ROTATED SPEECH

by

Barry A. Blesser

S.B., Massachusetts Institute of Technology
(1964)

S.M., Massachusetts Institute of Technology
(1965)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1969

Signature of Author_____
Department of Electrical Engineering

Certified by_____
Thesis Supervisor

Accepted by_____
Chairman, Departmental Committee on Graduate Students

PERCEPTION OF SPECTRALLY ROTATED SPEECH

by

Barry A. Blesser

Submitted to the Department of Electrical Engineering on June 1969 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

ABSTRACT

A new experimental technique, using spectral rotation, has been developed for studying the perception of speech in terms of its constituent parts. In the experiment described, pairs of subjects learned to converse with each other through a medium which transformed the high-frequency energy into low-frequency energy and vice versa. Spectral transformation, unlike distortion and filtering, creates a new sound pattern which is not initially intelligible but can be learned after about four hours of practice. An extensive series of tests were given at the end of each conversational practice session in order to gain insight into the perceptual relevance of phonemes, words, syntax, semantics, and prosodic features to the comprehension of speech. The tests measured discrimination of sound units, identification of phonemes, and the comprehension of words and sentences. In addition, the content of the practice sessions was analyzed for the kind of learning strategies utilized by the subjects.

The results of the tests showed that prosodic features, which are unaffected by the transformation, were more important than the phonetic cues for the comprehension of sentences. Subjects never learned to identify isolated sound units correctly; the perception of vowels seemed to be unstable, and the place of articulation feature for consonants was never learned. The relative ability of the subjects to perform discrimination, identification and comprehension tasks remained uncorrelated. The incorporation of meta-linguistic cues, such as

syntactic structure, semantic redundancy, and stress patterns, into cognitive strategies compensates for the loss of some phonetic cues.

The experiment suggests that speech cannot be reduced to a single set of necessary and sufficient cues. Moreover, comprehending a sentence is not simply the result of perceiving the words contained within it, nor are words perceived by identifying the constituent phonemes.

THESIS SUPERVISOR: Donald E. Troxel
TITLE: Associate Professor of Electrical Engineering

TABLE OF CONTENTS

FIGURES & GRAPHS

ACKNOWLEDGEMENT

The author wishes to thank his thesis supervisor, Professor D. E. Troxel, and readers, Professor M. Eden and Dr. P. A. Kolers for their interest and encouragement during the course of this research. Without their faith in the idea of being able to learn spectrally rotated speech, this experiment would never have come about. A special debt is owed to Dr. Kolers who acted as my mentor, providing me with a foundation for understanding psychological issues. The author also wishes to thank Phyllis Wilner for transcribing the data from the tape recorded conversations, Professor K. N. Stevens for helping me resolve some apparent contradictions, M. Lazarus for the initial conversation which led to this thesis, and Bess Themo for typing the final manuscript.

The facilities of the Cognitive Information Processing Group of the Research Laboratory for Electronics, whose work is supported in part by the National Institutes of Health and the Joint Services Electronics Program, were used throughout.

CHAPTER I

INTRODUCTION

The present interest in man-machine systems, in particular, machines which attempt to imitate human processes, has served as an added stimulus for investigating and modeling various cognitive functions. The difficulties encountered when the techniques of the physical sciences have been applied to such seemingly simple tasks as speech production, speech perception, and visual pattern identification have motivated much of the psychological research carried on by non-psychologists. Although engineering interest into the classical questions of psychology has sometimes shifted the criterion of an adequate explanation from "the ability to predict the outcome of an experiment" to the "the ability to improve or implement a system", one must not overlook, either implicitly or explicitly, the fundamental differences between a task-oriented machine and a conscious being. Thus, insight into cognitive perception may help the engineer but the nature of the investigation remains, and must remain, essentially psychological. The results of the research, however, do offer suggestions on how one might structure information processing machines (Kolers and Eden 1968).

During the initial stages of speech research it was assumed that a physical, in contrast to a perceptual, description of the acoustic wave would be a necessary and sufficient specification of the speech process. As the early optimism gave way to realism, the research focus became perceptual. Because of the enormous diversity of cognitive and perceptual problems, research has been segmented into a number of distinct orientations, including developmental linguistics, comparative linguistics, phonology, phonetics, semantic symbolism, syntactic structures, speech articulation, acoustical psychophysics, bi-lingualism, and synthetic speech. Each of these approaches has evolved a somewhat limited-scope definition of the problems it is willing to consider. Even by narrowing the range, however, most of the essential questions remain unanswered; and for this reason, more effort has been devoted

to working within a restricted range than in trying to integrate tenuous theories.

When considering the speech process, one must realize that it is a unity, and that a theory is made up of elements which are constructs--artificial constructs--whose value lies in their ability to be used to describe some aspect of the process. Description is not reality. A good example of this kind of situation is illustrated by the case of "distinctive features"*. Until recently, this construct was the providence of linguists who used it to describe the interrelation between the phonemic symbols of a language. However, the usefulness and sophistication of the construct has motivated a search for: the acoustic correlates of distinctive features, the articulation of distinctive features, the perceptual correlates of distinctive features, and, finally, the short-term memory of distinctive featues. The results of the search might demonstrate that the linguists had created a description that could be used in the other domains, but there is no a priori reason to expect this to be the case.

In the extensive investigations of speech perception, most researchers use phonemes or syllables in an isolated, context-free environment. Implicit in this approach is the view that a phoneme has a perceptual reality in natural speech. There is, however, no conclusive evidence to indicate that the way in which an isolated phoneme is perceived is the same as the way in which continuous speech is processed. The issue of physical and psychological reality is considered by Chomsky and Halle (1968, p. 25):

> "We take for granted, then, that phonetic representations describe a perceptual reality. Notice, however, that there is nothing to suggest that these phonetic representations also describe a physical or acoustical reality in any detail. For example, there is little reason to suppose that the perceived

*For a more complete discussion of distinctive features and phonetic cues see Section 4.1.

> stress contour must represent some physical property of the utterance in a point-to-point fashion; a speaker who utilizes the principle of transformational cycle and Compound and Nuclear Stress Rules should 'hear' the stress contour of the utterance that he perceives and understnads, whether or not it is physically present in any detail."

The reverse is also true; once having perceived the stress from the physical attributes of the acoustic wave, the listener can incorporate that information into his "Compound and Nuclear Stress Rules" in order to restrict the range of possible utterances which could be perceived.

The research described in this thesis is an attempt to demonstrate the multi-level nature of speech cognition by examining the perceptual behavior of subjects when listening to spectrally transformed speech. Thus, by creating a new "speech environment" rather than by limiting the normal environment, this experiment provides a global picture of the interaction and the relative importance of the perceptual levels. In one sense, spectrally transformed English speech is a new language which happens to have the same vocabulary and syntax as English but the actual sounds are alien or "foreign". Typical speech experiments are attempts to resolve very limited hypotheses concerning perception, such as how rapidly must a formant transition occur in order for the sound unit (isolated) be identified as a /b/ instead of a /w/. Or, at the other end of the scale, theoretical question concerning the syntactic rule may be investigated. However, in each case the investigation is conducted under very special condition with the stipulation that the results are valid under the assumption inherent in the experimental design. Spectral transformation is also a restricted condition, at the sound level, but it allows for the interaction of many cognitive levels. The nature of spectral transformation is considered in detail (see Chapter II) so that the relationship between normal English and transformed English may be determined. The conclusions from this experiment, in addition to suggesting a new perception model, emphasize the necessity of developing experimental tools

which can be used to explore the inter-dependence of semantics, syntax, phonetic cues, etc. Spectral transformation is one such tool.

1.1 THE SPEECH PERCEPTION PROBLEM*

Perception, as the activity which keeps the human organism in sensory contact with the physical environment, is the subjective inner sensation of a physical stimulus impinging on the sensory organs. Since the percept, in a naive behaviorist sense, is merely the response to a stimulus, the cognitive process may be viewed as that which extracts the "relevant" psychological information leaving the remainder as physical and psychological noise. By extracting the "essence" of the signal, the cognitive process equates signals which would otherwise not be equivalent. Thus, it is possible for a native speaker of English to recognize the sentence "I am going into the house." when spoken by any one of a thousand speakers; but, someone who did not understand English would be unable to make a judgment of similarity. In each case the acoustic wave is different. The phenomenon of speech perception, therefore, is a restatement of the question "How does the organism extract a single set of word symbols and meaning from unique acoustic signals?". Neisser (1966, p. 62) gives a very clear statement of the issue: "Without some definition or criterion of similarity no empirical prediction is possible; we are left to guess whether any particular stimulus will be recognized or not. Without any explicit model or mechanism, the notion of 'similarity' is only a restatement of the observed fact that some inputs are recognized while others are not."

During the past two decades, extensive experimentation on the acoustic characteristics of speech sounds has given researchers a sense that they are beginning to discover those attributes of the speech sounds which give rise to perception. As an example, let us consider a typical front vowel, /i/, and a typical back vowel, /u/. The /i/ sound contains a large amount of high frequencies energy in

*For a comprehensive bibliography covering recent speech research see Holmgren (1966).

contrast to the /u/ sound which is almost all low frequency energy. If one interprets this as meaning that these spectral attributes determine the perception of the vowels, then one must also say that subjects could not learn to understand spectrally transformed speech. Because the transformation reverses the high and low frequency energy, subject should interchange the perception of these two vowels. That is, the transformed /u/ should be perceived as an /i/ since it is more similar, acoustically, to the untransformed /i/ than to the untransformed /u/. Any restricted notion of perception as being cause-and-effect based on similarity must make this prediction. Yet, reality contradicts the prediction. Subjects do learn to understand spectrally transformed speech and they do not generally interchange the front and back vowels (after about 30 minutes of practice).

The difficulty with using physical similarity as a basis for perception, as Chomsky and Halle (1968, p. 294) point out, is that "even crude agreement between the external stimulus and the internally generated hypothesis suffices to confirm the latter. In other words the dependence of perception on properties physically present in the signal is less than total. What is more there are many extragrammatical factors that determine how close a fit between data and hypothesis is required for confirmation." Thus, perception is an active process whereby the listener "creates" the percept, and one of the tests for its acceptability is related to attributes of the acoustic stimulus. Such a theory is neither compact nor easy to verify, but it does have the major advantage of allowing for the possibility of understanding spectrally transformed speech. Moreover, there is the strong suggestion that there is no single set of attributes or cues which are needed for confirmation. Further insight into speech communications will come from discovering the other tests of acceptability and not from more accurate descriptions of the speech sounds.

1.2 DISTORTIONS AND TRANSFORMATIONS

When viewed as information transmission, speech communications is characterized by the encoding and decoding of word symbols in an

acoustic medium. The high degree of redundancy inherent in the modulation is exemplified by the fact that the information rate of the phonemic sequence is on the order of 50 bits/second and the information rate of the acoustic wave is as much as 30,000 bits/second (Flanagan 1965, p. 4). In a mathematical sense, the relevant information is distributed throughout the speech wave; and only a small undistorted portion of the original wave is usually sufficient for intelligibility. "Beginners in the study of privacy systems never fail to be amazed at the difficulty of scrambling speech sufficiently to destroy intelligence. The ear can tolerate or even ignore surprising amounts of noise, non-linearity, frequency distortion, misplaced components, superposition, and other forms of interference" (Kahn 1967, p. 588). "The fact that the ear is such a good decoding tool makes the production of privacy systems very difficult. Scrambling systems which look effective on paper sometimes turn out on trial to degrade the intelligibility very little, although scrambled speech usually sounds unpleasant" (ibid. p. 599).

Rather than looking for the "essence" of the intelligibility, one can take the alternative approach, namely, discarding aspects of the signal which do not make a significant contribution to intelligibility. Basically, if the gross spectral and temporal pattern of the speech is preserved then it will be intelligible. Perhaps the best illustration of this comes from the work with vocoders, which are systems designed to transmit bandwidth compressed speech. By sending only the slowly varying spectrum of the speech instead of the actual signal, these systems can typically operate at an information rate of 2400 bits/second (Voiers 1968) and, under special condition, at rates as low as 900 bits/second. The fact that a limited set of parameters describing the spectral shape of the speech is sufficient to transmit an intelligible message has sometimes been interpreted as meaning that the spectrum is a first step to finding an "essence". Also, it is generally believed that the spectrograph, or time-frequency-amplitude, description of speech portrays the important aspects of the signal. In fact, extensive effort has been directed at trying to learn to "read"

spectrographs (Potter, Kopp and Kopp 1966), but these efforts have met with only partial success.

The amount of spectral information required for a given intelligibility level has been extensively investigated by French and Steinberg (1947). They created a method for calculating the articulation index as a function of intensity of the frequency bands and the relative signal-to-noise ratio. Licklider, Bindra, and Pollack (1948) demonstrated that peak-clipped speech, in contrast to center-clipped speech, is intelligible because the spectral pattern of the original speech signal remains intact. Speech synthesizers, using spectral parameters to control the generated speech, can produce very intelligibile speech. Confirmation of the theroy that the spectrum is both necessary and sufficient is found in this experiment with transformed speech. Naive subjects cannot understand speech when the spectrum is rotated.

In this experiment, the spectrum of the speech signal was rotated so that the high-frequency energy became low frequency energy and vice versa. This kind of transformation has the property that none of the information, from an information theory point of view, is removed; it is merely re-encoded into the acoustic signal. Such a transformation, in contrast to additive noise, distortion, and filtering, does not destroy any of the original information. Rather, it creates a new signal which bears a one-to-one correspondence to the old signal. Moreover, it is a transformation that is singularly appropriate since the preliminary auditory processing appears to decompose the incoming acoustic wave into spectral components.

Although a distorted speech signal and a spectrally transformed speech signal may be equally unintelligible initially, subjects can learn to understand the transformed speech after extensive practice. This contrasts with filtered or distorted speech experiments in which very little learning takes place regardless of how much practice is given. Thus, the processes involved in the two cases are very different: transformation is a very special kind of distortion. Even though

cognitive processing of speech is far subtler than mere feature extraction from the spectral pattern, some aspects of the spectral pattern contribute to perception. In the transformed speech case, these aspects are not available to the listener initially but after some exposure to the new medium they again become sufficiently available to be used. Just what these aspects or features are remains a moot point. The experienced listener does not simply "re-transform" the pattern.

The original motivation for this experiment came from the visual transformation investigations of Stratton (1897), Kohler (1964) and Held (1965). By wearing prisims which rotated, shifted or inverted the visual pattern, subjects experienced a position or orientation change in their visual field. Like the spectral transformation, the visual transformations are characterized by a one-to-one correspondence between the new and old patterns and by a constancy in the relative distance between two points in the field. In the visual experiments, subjects initially experienced a sense of disorientation, but they soon adapted so that their visual sense of reality became natural.

Because subjects can learn to adapt to both the visual and auditory transformation, one cannot view perception as a mechanized process of feature detection using a fixed set of "wired-in" computation sub-systems.

1.3 OVERVIEW

The performance of subjects when listening to spectrally transformed speech was first tested in a pilot study using one pair of subjects. They were placed in the transformed medium and given the task of learning to understand each other. Although these two subjects later proved to have been unusually facile in their ability to learn to communicate with each other, the fact that they did not understand each other initially and that they were able to converse after only about 2 hours of practice served as the foundation of the experiment described herein.

In the main experiment, an extensive set of tests, which attempted to measure changes in the subjects' ability to perform such tasks as discriminating sounds, identifying phonemes, recognizing words, and comprehending sentences, was given at the end of each session. The results of the tests try to answer the question "What have the subjects learned as they improved their ability to converse with their partner?" If, for example, comprehension is related to identifying phonemes, then one would expect that those subjects who did well on the phoneme tests should also have done well on the comprehension tests; or, when subjects could no longer improve their performance on identifying phonemes, then their ability to understand sentences should also have reached a limit. Actually, the situation was exactly the opposite. The psychophysical discussions in Chapter II and the elaborate description of the phonetic and phonemic performance in Chapter IV are included mainly for the purpose of demonstrating that competence in understanding conversational transformed speech is almost independent of competence on isolated speech tasks. Even identification of words, which are far more complex than phonemes, appears to be almost unrelated to sentence comprehension, as discussed in Chapter V. The data from the tests have been examined in a variety of ways in order to show this; the correlation studies and the learning curves both show that caution is needed when one tries to relate aspects of the speech process to the whole. In other words, speech perception occurs at a level which is so much more complex than feature extraction, that features, as they are presently understood, often appear to be almost unrelated, in the case of sentences, to the percept. This, actually, forms the underlying theme throughout most of the experiment.

Under ordinary conditions, little mental energy is needed for speech communications and the cognitive mechanisms are freed from the burden of concentrating on the act of perception. Because normal speech perception is so rapid and automatic, it is very difficult to gain insight into the process. With spectrally transformed speech, subjects must re-establish automatic associations between new sound patterns and verbal symbols. Analysis of the conversations in Chapter VI suggests that this is in fact what the subjects were learning, and

that this learning is facilitated by the integrative prosodic patterns and not by the phonetic features. In a sense, there is a parallel between learning to understand spectrally transformed speech and learning a foreign language or a first language. The four stages of learning, as discussed in Chapter VI, might thus be a reflection of the perceptual process itself.

The fact that speech is an integral part of a personality and not an independent subroutine is often overlooked when one examines it as an isolated function. The quality and character of the conversation varied greatly, illustrating that the subjects brought their entire personality, as well as their language competence, into play when they were trying to understand transformed speech. This phenomenon manifests itself in the kind of strategies used by the subjects. Moreover, the meta-language factors which are an inseparable part of speech communication often influenced their ability to learn the new medium.

Since the experiment attempted to treat the entire speech process in the context of spectral distortion, the results cannot be integrated into one complete theory. Cognition is not that well understood. Rather, the results should be viewed as a strong suggestion about how one ought to view the process and, in this sense, it points to a new kind of experimentation for exploring speech.

CHAPTER II

CHARACTERISTICS OF TRANSFORMATION

It is impossible to imagine a sound which has been spectrally inverted since the transformation is not a naturally occurring phenomenon. Only some communications engineers and radio operators familiar with single-side band reception come into contact with such sounds, whereas, the transformations in the visual field produced by mirrors and lenses are quire commonplace. If one tries to imagine how a spectrally rotated sound would be perceived, one might create an acoustic image in which the melody of a sound was inverted or in which a question was changed to a statement. In fact, however, the transformation does not affect the sense of pitch, instead, it affects the "quality" of the sound. Correlating the sensation of a sound with the spectrum is a rare ability, although some spectrally related concepts do exist, as illustrated by a musician describing a given musical tone as "harsh" or "melodic". These two adjectives describe aspects of the spectrum which would be changed by the transformation. Thus, it is important to consider the characteristics of the transformation from both a physical and a psychological point of view.

2.1 MODEL OF AUDITORY PATTERNS

The effect of the spectral rotation transformation on speech is clarified by viewing the act of perception as pattern recognition. Allegedly, the original acoustic wave is continuously being processed, stage-by-stage, until it is reduced to a set of features or characteristics. Another way of saying this is that at any stage the neural firing rates are a pattern which results from the processing of the pattern in the previous stage. The features are therefore characterized, according to this abstract theory, as the pattern which exists in the central nervous system. Actually, almost nothing is known about the nature of neural patterns for complex signals at any stage of processing.

Present knowledge about the peripheral stages of auditory processing suggest that the neural pattern in the auditory nerve leaving the hair cells is not similar to the original acoustic wave; rather, the two-dimensional amplitude-time wave is transformed by the cochlea into a three dimensional amplitude-space-time pattern. The mechanical resonant properties of the Basilar membrane combined with neural inhibitory contrast, known as Mach's law of contrast (Ratliff 1961), result in each of the 30,000 nerve fibers of the auditory nerve being tuned to a different frequency component of the incoming acoustic wave. Physiological evidence of this tuning has been shown for cats (Kiang 1965). Thus, the space dimension, or position along the basilar membrane, is a monotonic function of frequency. The central nervous system performs, in a way not presently understood, a pattern recognition of the firing rates as a function of time for the entire nerve bundle.

The schematic of the model just described is shown in Fig. 2.1a (Siebert 1968). From a mathematical point, the generality of the model is so great that a given firing pattern of the auditory nerve bundle over-specifies the input which generates it. Thus, one is not tempted to use such a model to determine which aspect of the signal is important to perception and which can be discarded. Rather, the power of the model lies in its suggestion that spectral rotation of the acoustic wave transforms only the space part of the three dimensional pattern, leaving the dimensions of time and amplitude unaffected. The transformation follows a rather simple algorithm: the time firing pattern of a nerve which would have responded to a low frequency component before transformation now responds to a high frequency component. This is equivalent, in our simple model of auditory processing, to cutting the auditory nerve and re-splicing it with different parts of the cochlea exciting new parts of the higher auditory center.

Moreover, Siebert's model also suggests that the transformation is more than a distortion of the acoustic signal. Consider, for example, distortion introduced by hard limiting. A speech signal which has been hard limited so that the time waveform is either 1 or

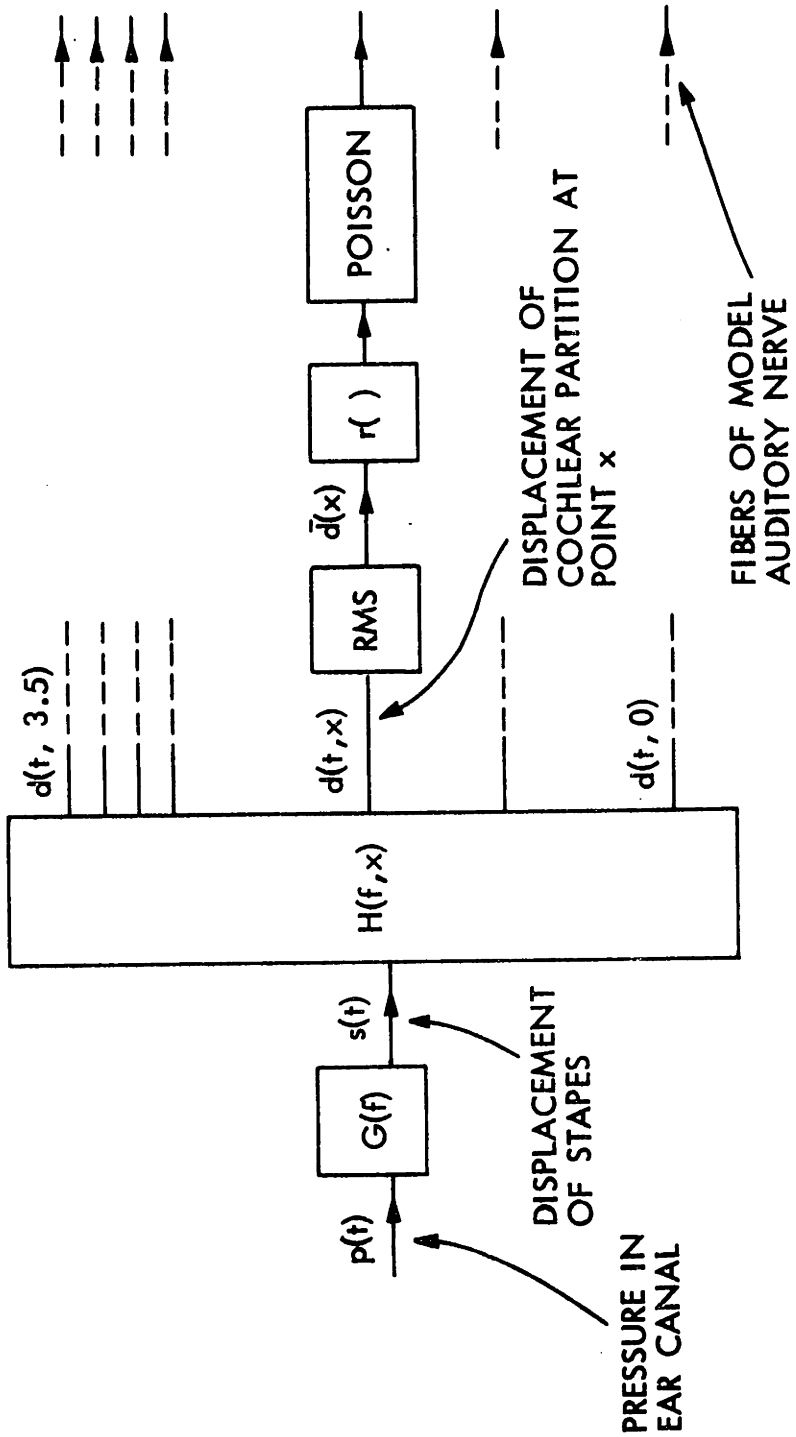


FIG. 2.1a MODEL OF PERIPHERAL AUDITORY SYSTEM (SIEBERT 1968).

-1 appears to have very little of the original speech signal remaining when viewed in the time domain. However, since the energy density spectrum of the hard-limited speech has the same frequency components as the original, in addition to those generated by the distortion, it is argued that the original neural firing pattern of the hair cells is still present even if somewhat obscured by the presence of the distortion-related components (Licklider, Bindra, and Pollack 1948)*. Perceiving a pattern which has resulted from distortion is analogous to detecting a desired signal in the presence of noise; whereas recognizing a new pattern is felt to be basically a different problem. Confirmation of this notion is found in the fact that spectral rotated speech can be learned, while distortion degrades intelligibility to a relatively fixed level.

Unfortunately, the linear frequency transformation does not produce a linear space transformation. The relationship between the maximum position of stimulation along the cochlea and the resonant frequency, according to the data of Békésy (1960, Chs. 11 and 12), is approximated by

$$x_o(f) \ln 10 = \ln \frac{10^5}{f}$$

where f is measured in Hz and $x_o(f)$ is the distance from the stapes in cm. Since the nerve density along the basilar membrane is approximately linear, the effect of the transformation is to change the nerve density as a function of frequency. For example, components of the acoustic signal between 300 and 400 Hz , which stimulate some number of fibers before transformation, become components between 2800 and 2900 Hz. (assuming 1600 Hz rotation frequency) and now stimulate a different number of fibers. Since the surface of the basilar membrane is somewhat analogous to the skin in a tactile sense, and since two-point

*Thomas (1968) showed that hard-limiting the second formant produced intelligible speech, but hard-limiting the first formant did not.

thresholds* on the surface of ths skin are proportional to nerve density, one might suppose that the transformation would change the auditory sensitivity to a particular component (Békésy 1960, p. 566). This differential sensitivity is unavoidable, because it is virtually impossible to implement a frequency transformation which would produce a linear pattern transformation using a real-time system.

Another implication of Siebert's model is that the short-term average intensity of the spectral components controls the mean firing rate of the nerve cells. This averaging destroys the phase relationships between the various frequency components. Although various experiments have shown the existence of time synchronization, or phase locking, for low frequency signals, and to a lesser extent for medium frequency components, Goldstein (1959) contends that any detailed time structure varying faster than 200 Hz is not used by the more central parts of the nervous system. This is quite fortunate, if true, since one of the effects of the transformation, as shown in the next section, is to produce an output signal whose phase relationships are time varying.

The optical analog of the auditory transformation is the reversal or rotation of the visual field. In the experiments of Stratton (1897) and Kohler (1964) either the horizontal or vertical dimension was reversed by using inverting prisms. The major difference between the experiments with the two sensory modes is that vision has two spatial dimensions and audition has only one, frequency.

2.2 MATHEMATICS OF THE SYSTEM

The effect of the spectral rotation system on the incoming signal is impossible to visualize if the signal is represented as a time-varying intensity, either electrical voltage or acoustic pressure.

*Two-point threshold is the minimum distance between two point-pressure stimuli which can be detected; below the threshold the two stimuli appear as one. The threshold is directly related to the nerve density, i.e. the more nerve cells leaving the skin surface the lower the threshold.

Every time-domain representation, specifying an intensity as a function of time, can be rewritten as a frequency-domain form, specifying an intensity as a function of frequency (see e.g. Lathi 1965). The two functions form a Fourier transform pair and are equivalent since one can be generated from the other.

Let $F_o(w)$ be the frequency representation of $x_o(t)$, the output signal from the transform system, and let $F_i(w)$ be the frequency representation of $x_i(t)$, the input to the system. The relationship between these two functionsis

$$F_o(w) = \begin{cases} e^{j\theta} F_i (2w_c - w), & 2w_c > w > 0 \\ e^{-j\theta} F_i (-2w_c - w), & -2w_c < w < 0 \end{cases} \tag{1}$$

where w is the frequency in radians, $w_c/2$ is the center frequency in Hz about which the rotation takes place, and θ is an arbitrary phase angle dependent on the initial phase at t = 0. Perhaps, in order to fully appreciate why the system must be viewed in terms of the frequency-domain, consider the time-domain version of (1) as

$$x(t) = e^{j\theta}\int_{o}^{+2w_c}\left[\int_{-\infty}^{+\infty} x_i(\tau)e^{-j(2w_c-w)\tau}\, d\tau\right]e^{jwt}\, \frac{dw}{2\pi}$$

$$e^{-j\theta}\int_{-2w_c}^{o}\left[\int_{-\infty}^{\infty} x_i(\tau)e^{+j(2w_c+w)\tau}\, d\tau\right]e^{jwt}\, \frac{dw}{2\pi} \tag{2}$$

This integral equation cannot, in general, be solved except under special restricted conditions. Solving it for some particular input requires that one, in effect, go through the steps of first finding the Fourier transform representation of the input signal, rotating the frequency function, and then finding the inverse Fourier transform.

To illustrate the operation of the spectral rotation system,

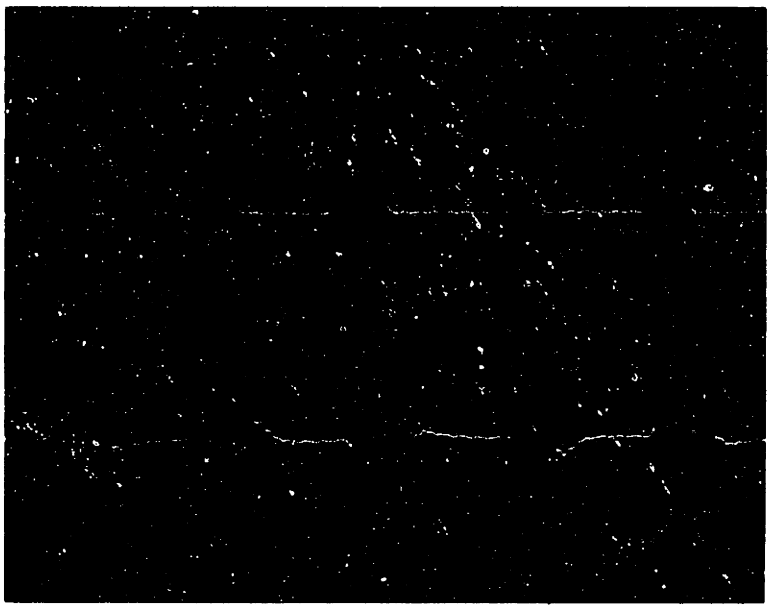
first consider a sinusoidal input at 300 Hz. With the 1600 Hz center frequency used in this experiment, the output is also a sinusoidal signal, but at a frequency of 2900 Hz. Likewise, an input at 1000 Hz produces an output at 2200 Hz. The output signal with a more complex periodic signal, as for example a square wave, can easily be found by using the property of linearity. One can show, using either equation (1) or (2) that with an input of $\alpha x_i + \beta y_i$ the output is $\alpha x_o + \beta y_o$, if x_o and y_o are the outputs with inputs x_i and y_i respectively. Since this is a necessary and sufficient condition for linearity, the output with a periodic input is simply the sum of the outputs produced by each of the input components individually. A 300 Hz square wave contains energy at frequencies of 300, 900, 1500, 2100, and 2700 Hz. Using equation (1), which in its simple form is written as

$$F_o(w) = F_i(2w_c - w) \tag{3}$$

one can see that each component is rotated about 1600 Hz to produce an output which contains components at 2900, 2300, 1700, 1100, and 500 Hz. The intensity of magnitude of the output harmonics is equal to the intensity of the input harmonic which produced that part of the output. Although the 300 Hz square wave contains harmonics above 2700 Hz, these are filtered out to prevent spurious components from appearing. The low-pass filtering at 3200 Hz removes components which are greater than twice the center rotation frequency*.

The above example also illustrates that, in general, true periodicity is destroyed since all the components at the output form an anharmonic series and are not multiples of the repetition rate. The true repetition rate at the output with a periodic input is the lowest common factor in each of the harmonics. This effect is demonstrated by using a 1600 Hz tone burst sequence at a repetition rate of about 350 Hz. Although the output, as shown in Fig. 2.2a, is also a tone burst

*The actual filter used in the system removes all components above 3000 Hz and below 200 Hz, thereby creating a 200 Hz guard band around the transformation.



(A)

(B)

FIG. 2.2a SEQUENCE OF TONE BURSTS AT 1600 HZ.

(A) input.

(B) output.

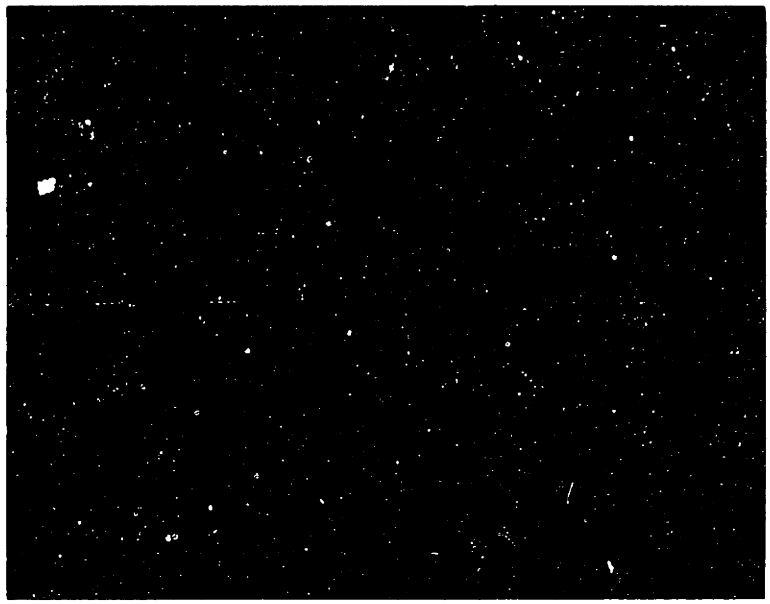
sequence at the same frequency, it is not periodic since it no longer satisfies the constraint.

$$x(t-T) = x(t) \tag{4}$$

where T is the period. This phenomenon can also be explained in terms of the random **phase angle**, θ, in equations (1) and (2). At the beginning of each pulse, the phase angle is essentially random. Because the components of the signal have a new phase relationship, the wave shape is different. If the phase factor were proportional to frequency, the effect would be simply to delay the pulse, but it is constant. Thus, the same input can produce an infinite variety of outputs all of which have the same energy density spectrum but a different phase function. Only periodic signals with a period equal to a multiple of π/w_c produce truly periodic outputs.

Although the output from the system with periodic input signals is easy to describe, **truly** periodic signals never occur in speech. At best, a voiced phoneme may appear to satisfy the definition of (4) for as much as 100 milliseconds. It can be shown, however, using a time limited version of the Fourier representation (Fano 1950) in equation (1), that the spectrum of a segment of the output waveform is the spectral transform of the input spectrum for the same time segment. Thus, the system produces a quasi-periodic output for a quasi-periodic input (in the same sense that a periodic input produces a periodic output). This is illustrated in Fig. 2.2b which shows the outputs for two tone bursts at different frequencies. The 2200 Hz tone burst becomes a 1000 Hz tone burst.

To consider the effect of the system on speech, let us temporarily assume that all speech signals are composed of either quasi-periodic vocal sounds, noise burst fricatives, or a combination of both. The voiced phonemes contain a distinctive repetition rate or pitch and a spectral coloring which distinguishes one phoneme from another. The spectrum of a typical vowel, /e/, before and after spectral transformation is shown in Fig. 2.2c. In both cases, the



(A)

(B)

(C)

(D)

FIG. 2.2b EFFECTS OF SYSTEM ON TONE BURSTS.

(A) input at 700 Hz.
(B) output for above.
(C) input at 2200 Hz.
(D) output for above.

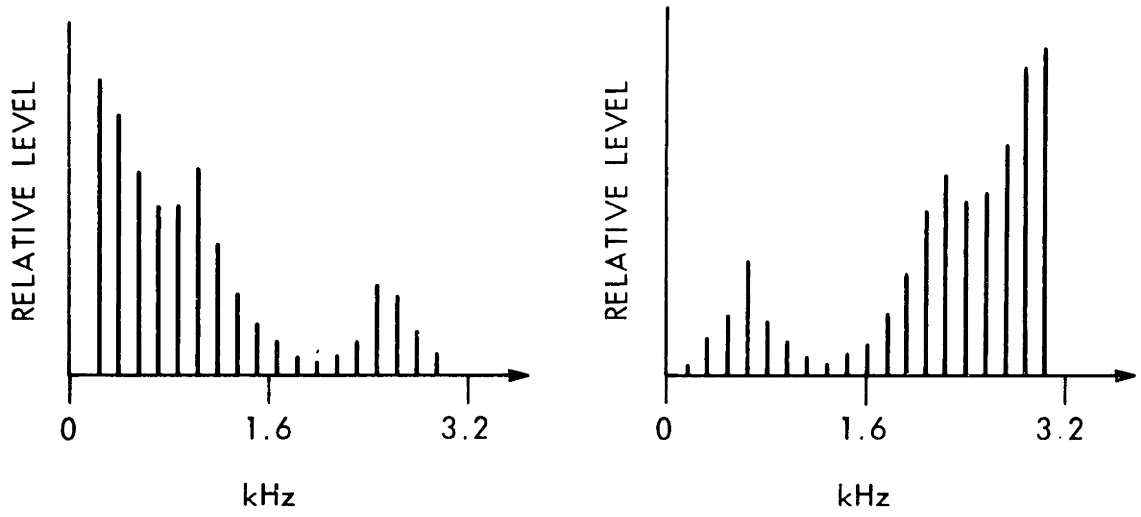


FIG. 2.2c SPECTROGRAPH RECORDING OF /e/ BEFORE TRANSFORMATION (left) AND AFTER TRANSFORMATION (right).

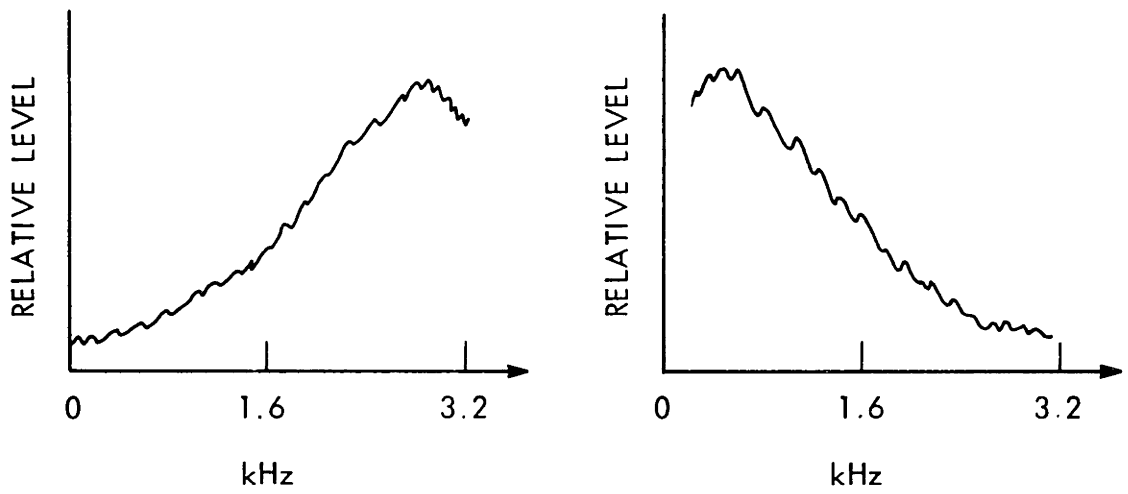


FIG. 2.2d SPECTROGRAPH RECORDING OF /ʃ/ BEFORE TRANSFORMATION (left) AND AFTER TRANSFORMATION (right).

subjective pitch, equal to the distance between harmonic components, is the same; but, the spectral envelope that gives the vowel its distinctive sound is rotated about the 1600 Hz center frequency. Although the location of the first component of the output is equal to the smallest positive difference between 3200 Hz and the highest harmonics of the original, the subjective pitch is unaffected by this value.

For a steady-state fricative, such as /s/, the spectrum is continuous with no distinct components. This phoneme, as shown in Fig. 2.2d, is simply colored noise. When the signal is transformed, the output is also colored noise, but with a different coloring. The spectrum of the output is simply the input spectrum rotated about the center frequency. There is no question of periodicity, harmonics, or pitch.

2.3 PSYCHOPHYSICS OF TRANSFORMATION

If the ear were equally sensitive to all frequency components, the sole effect of the spectral transformation would be the creation of a new pattern. It was shown in Section 2.1 that the density of nerve fibers is not constant or independent of frequency. Furthermore, there is no assurance that the neural interconnections perform the same functions over the entire frequency range. The present knowledge about the kinds of complex processing that must exist in the auditory nervous system is so meager, that, at best, differential sensitivity and performance can be assessed only with simple stimuli. The fact that even in the region from 200 to 3000 Hz the performance of the auditory system to simple psychophysical tasks is dependent of frequency means that, in additon to learning to perceive an unfamiliar pattern, the auditory system has a different sensitivity to the new pattern. As an obvious example consider what would happen if the speech band were shifted up to 5000 Hz. The listener would never learn to perceive these new patterns simply because the auditory system does not perform well at very high frequencies. One measure of this effect is the differential performance to psychophysical tests as a function of frequency.

The ability of subjects to discriminate small changes in amplitude and frequency has been investigated by many researchers under different conditions. In an experiment using a single tone warbled at a 2 Hz rate, Shower and Biddulp (1931) showed that in the region from 150 to 1000 Hz the difference limen for frequency was approximately constant. This result contrasts with the test of Rosenblith and Stevens (1953) using an ABX method; they found that the JND value at 1000 Hz was more than twice the value at 250 Hz.

Differential sensitivities in other tasks suggest that there are some basic differences in perception above and below 1000 Hz. The resolving time needed for judging a given frequency is about 2 or 3 Hz for frequencies below 1000 Hz, but relatively constant above this value (Doughty and Garner 1947). The function which relates subjective pitch to frequency is linear below 1000 Hz and logarithmic above (Stevens and Volkman 1940). The intensity of noise required to mask a single tone also approximates the curve for subjective pitch (Fletcher 1938). The critical frequency bands used in loudness calculations show the same trend (Zwicker, Flottorp, and Stevens 1957), as do the articulation bands used in speech intelligibility calculations (French and Steinberg 1947). Although the difference limen for amplitude changes is independent of frequency when measured at a constant level above threshold, the threshold level is constant from 1000 to 3000 Hz, but increased for low frequencies (Békésy 1960, p. 217). Koenig (1949), using a frequency scale which was linear up to 1000 Hz and then logarithmic, found that most psychophysical data for hearing produced a linear curve.

Electro-physiological experiments with the cochlea in cats suggest that the difference in sensitivity is a reflection of different modes of neural encoding. For very low frequencies the mechanical resonance of the basilar membrane is not very pronounced and the firing patterns of the nerve fibers is in synchronization with incoming acoustic signal. At higher frequencies above 1000 Hz, the nerve cell response is not phase locked with the signal, rather it only gives a measure of the intensity at a particular frequency (Galambos 1958).

The response for medium frequencies is a combination of time and frequency. Furthermore, the shift in resonance position along the basilar membrane for a given change in frequency is approximately constant from 300 to 1000 Hz, whereas above this value it is proportional to a percentage change in frequency (Békésy 1944). These findings coincide with the sensitivity measurements.

In order for the spectral transformation to create a new pattern which is, from a neurophysiological view, equally easy to recognize, it would be desirable to work in the region from 200 to 1000 Hz. Unfortunately, this is not nearly sufficient for speech; the syllabic articulation factor for speech bandlimited to 1000 Hz is only 35% (French and Steinberg 1947). Thus, even without spectral transformation, this speech would be difficult or impossible to understand. One must chose a frequency band which is as small as possible to avoid the frequency variation of the auditory system, yet the band must be large enough so that speech is easily understood without spectral transformation. The band from 200 to 3000 Hz with a 1600 Hz center rotation frequency is a good compromise since it has a syllabic articulation index of 90%.

One finds that many of the fricative phonemes, as for example /f/ and /s/, are confused by bandlimiting speech to 3000 Hz since their distinguishing characteristics occur in the high frequency region. In isolation these phonemes are pure noise, and only the spectral coloring can be used to identify them. Thus, following the spectral transformation they are still confused.

The transformation causes the spectral components which used to lie in the more sensitive 200 to 1000 Hz region to be shifted to the less sensitive 2200 to 3000 region. And correspondingly, the part of the spectrum that was in the less sensitive region where the nerve firing pattern did not contain any synchronous information now excites the nerves in a temporal as well as spectral mode. In terms of the auditory threshold, the situation is exactly reversed since the threshold at 200 Hz is as much as 25 dB higher than that at 3000 Hz. The

implication of the changes in threshold and modes of firing can not be evaluated from the existing knowledge of the nervous system. It may be said, however, that spectral transformation of the audio signal results in a new pattern which differs perceptually from the old one in some complex manner.

2.4 PERCEPTION WITH COMPLEX SIGNALS

The spectral transformation, as mentioned in Section 2.3, creates a new pattern in the space, or frequency, dimension which is not a simple transformation of the nerve fiber firing patterns. This is illustrated, for example, by the fact that masking is not unilateral. An intense frequency component of the signal has a much greater tendency to mask higher frequencies than it does lower. In the presence of a 1200 Hz tone at 80 dB spl*, an 800 Hz tone is masked by about 10 dB whereas a 1600 Hz tone is masked by about 45 dB (Wegel and Lane 1924). The calculation techniques used to measure subjective loudness of complex signals includes an involved algorithm to account for the amplitude and frequency dependent masking (Zwicker 1958; Feldtkeller and Zwicker 1956). After spectral transformation, some harmonics which were previously masked may now be perceived and other components which were previously perceived may now be masked.

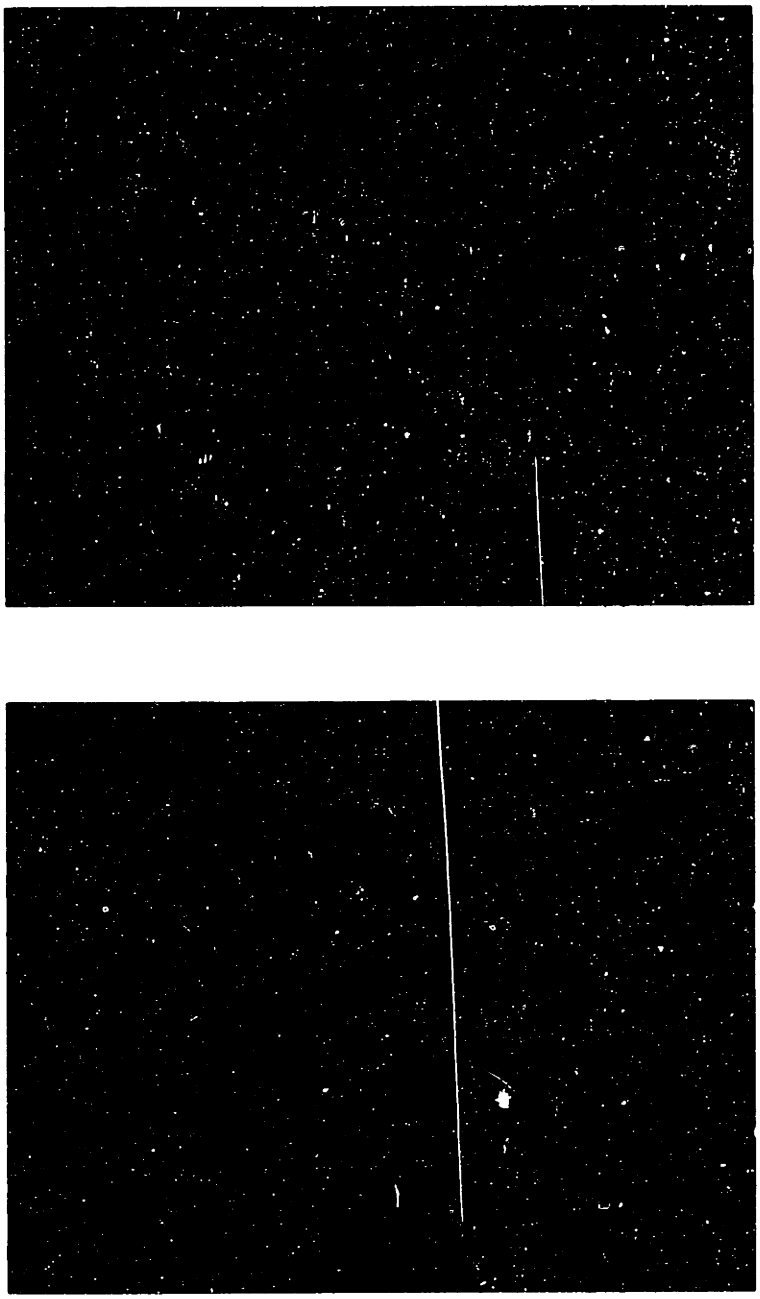
Experiments have shown that a signal composed of multiple tones or harmonics has a tendency to create beat frequencies, or the sensation of frequencies, equal to the difference between the component frequencies. In the case of an untransformed periodic sound, these new beat frequencies are equal to, or multiples of, the existing components. However, for the transformed sound, these beat frequency components are the same as those of the original untransformed sound. Thus, one might argue that some of the original pattern remains after spectral transformation.

*dB spl is sound pressure level with the 0 dB reference equal to .0002 dynes/cm^2.

Perhaps, an important difference between the transformed and untransformed sound is that the latter is never periodic, whereas the former is. A periodic signal whose harmonics are all multiples of the fundamental repetition rate is transformed into a signal whose components are not multiples of any fundamental frequency. Although the sensation of pitch is determined by the fundamental frequency, subjective pitch of the transformed sound does not seem to be affected. "Mathes and Miller showed that subjective pitch usually corresponds to the envelope-modulation frequency if the modulation envelope is at all pronounced" (Licklider 1956). A time waveform for the two vowels, /a/ and /u/, before and after transformation is shown in Fig. 2.4a. There is a very distinct envelope corresponding to the original pitch in the transformed case, even though there is no periodicity in the sense that each of the original segments of a single period is changed into a variety of segments.

deBoer demonstrated that with a sound composed of an inharmonic series the sensation of pitch is approximately equal to the difference between the harmonics (Plomb 1967). Furthermore, he showed that the fourth and higher harmonics determine the pitch for repetition rates up to 350 Hz (periodicity). With speech or any quasi-periodic signal the difference between the harmonics after transformation is equal to the original pitch. Experimentally, subjects found that pitch was not affected by the system. Although they initially expected to find a rising pitch to be turned into a falling pitch, and a question to be turned into a statement, this did not occur. Only whistling, which is almost a pure tone, illustrates a pitch transformation.

Unlike pitch, loudness is changed by the spectral rotation. The natural acoustic environment in which we live contains sounds whose energy density spectrum has a much greater value at low frequencies. After transformation this situation is reversed, and the sound quality is unpleasant and unnatural. Low frequency noise which is often present but unnoticed in the background becomes a 3000 Hz whine. The octave between 200 and 400 Hz which can contain a clearly



(A)

(B)

(C)

(D)

FIG. 2.4a TIME WAVEFORMS FOR VOWELS.

(A) /u/ before transformation.
(B) /u/ after transformation.
(C) /a/ before transformation.
(D) /a/ after transformation.

observable rhythm is transformed into a fraction of an octave from 2800 to 3000 Hz. The new sound appears to be a steady drone rather than melodic.

The low frequency first formant of speech contains much more energy than the other formants, yet it carries much less information. Before transformation, the first formant stimulates a region in the cochlea which has a high threshold; the more important low energy second formant occurs in a region of maximal sensitivity. Following transformation this situation is reversed. To prevent the first formant, which can be as much as 20 dB more intense than the second formant, from dominating other components, an equalizer shown in Fig. 2.4b was added. The equalizer accentuates the high frequencies and attenuates the low frequencies so that masking is reduced and the speech sounds more pleasant.

The long-term average energy density spectrum for speech is shown in Fig. 2.4b along with the spectrum at the output of the system with and without the equalizer. Casual tests showed that the "equality" of speech sounds was not significantly affected by the equalizer, and subjects no longer complained about the high-frequency drone.

With the equalizer the effective threshold for a given frequency component remains unchanged. A component at 200 Hz, for example, has a threshold which is about 25 dB greater than that at 3000 Hz. (Fletcher and Munsen 1933). After transformation the component formerly at 200 hz appears at 3000 Hz but the equalizer has reduced its amplitude by about 20 dB, thus restoring the threshold. At higher levels around 80 dB spl, the ear is equally sensitive to tones from about 200 to 2000 Hz, but with the equalizer this is no longer true. However, it was felt that this effect was less important than the advantages of having the equalizer.

2.5 SPECTRALLY INDEPENDENT FEATURES

According to our present understanding of the acoustic cues and

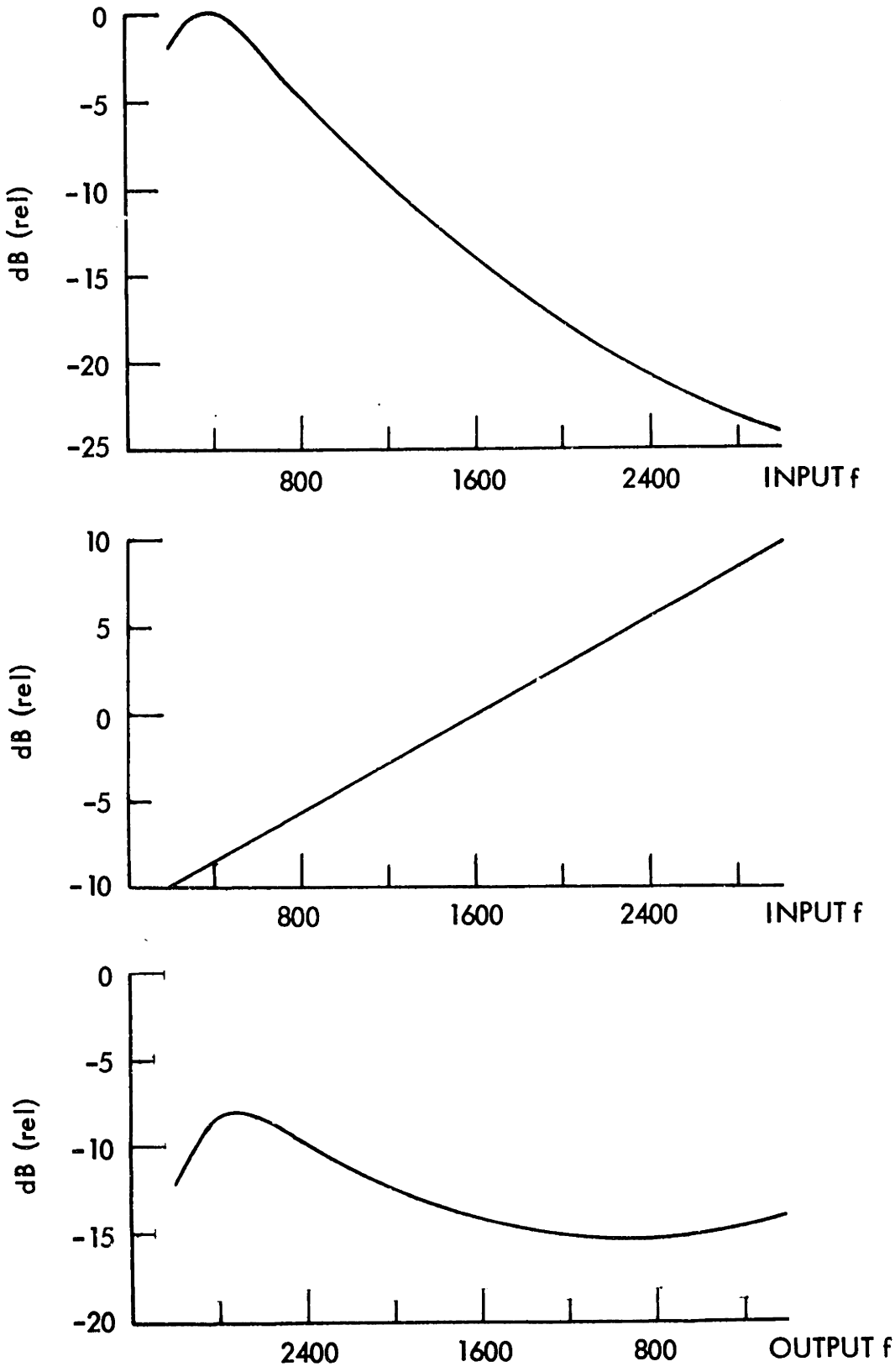


FIG. 2.4b LONG-TERM AVERAGE ENERGY DENSITY SPECTRUM FOR SPEECH.

(A) untransformed speech (after French and Steinberg 1947).

(B) frequency characteristics of equalizer.

(C) spectrally transformed speech.

features which play a role in the perception of speech, there are two classes: spectral and nonspectral. Because the features in the second category are unaffected by the transformation, they can be understood without any need for learning and may be a significant factor in the learning of the spectral cues. One finds within this class most of the parameters thought to be responsbile for the prosodic features, as well as many of the cues which enable the listener to distinguish individual phonemes. There are four sets of features that are independent of the transformation: pitch, loudness, temporal order, and the binary source features, voicing-nonvoicing, plosive-nonplosive, and fricative-nonfricative.

It was stated in the previous section that the inharmonic series resulting from transformation of the quasi-periodic voiced phonemes is perceived as having a subjective pitch equal to the original vocal repetition rate. Thus, the perceived pitch contour of an utterance is the same as the untransformed original. The rising pitch at the end of an utterance, characteristic of a question, is reproduced, rather than being inverted; a question remains a question and a statement remains a statement. The higher pitch for stressed syllables and words is also unaffected so that the accent is on the appropriate syllable. It may be argued that the spectral envelop and fundamental pitch are extracted as separate features since the perception of vowels is independent of the pitch. The fact that pitch is unaffected by the transformation is felt to be the explanation for such curious phenomenon as perceiving a sentence which differs from the spoken sentence in every way except that the part of speech for each word is correct (see Section 5.31).

As a phonetic cue, the mere presence of a periodic component indicates the existence of a voiced phoneme, as opposed to its unvoiced counterpart, e.g. the phoneme pair /f/ and /v/*. Similarly,

*The phonetic cue for distinguishing voiced and unvoiced may actually be unrelated to voicing and periodicity. The final /s/ is sometimes distinguished from a /z/ by a duration cue.

the noise component of a fricative phoneme is still a useable cue for the phoneme since noise is transformed to noise. The sudden release of energy characteristic of the plosive is reproduced after transformation, leaving the essential feature unchanged. Many of the binary distinctive features which distinguish one class of phonemes from another are not a function of the spectral envelope. The results of the phoneme identification tests, found in Sections 4.2 and 4.3, show that most confusions are within one class. That is, a /b/, /d/, and /g/ are much more readily confused since they are all voiced plosives; whereas, a /b/, /p/, /f/, /v/ and /m/ are all members of different classes and, therefore, easier to distinguish.

Intensity is also relatively unaffected by the system. Since the area under the input and output energy density spectra are equal to each other and proportional to intensity, the input and output signals have the same short-time average intensity. Loudness, however, being a perceptual measure, is somewhat changed by both the spectral transformation and the frequency equalizer. Nevertheless, for a particular sound with a specific spectrum, the output loudness is proportional to the input loudness. Thus, perceptability of the stress contour is further enhanced as is the distinguishability of the low and high energy phonemes.

The temporal order of the sound pattern is perserved by the transformation since the system is working in real-time. The duration of a phoneme remains the same so that a stressed tense vowel, being longer than its lax counterpart, retains its significant feature. A pause is mapped directly into a pause of the same duration since transform of a silent interval is still a silent interval. This allows the listener to identify a plosive by the preceeding stop, and it also helps in the perception of syntax-related juncture.

The sequential order of the phonemes in an utterance is perserved. The vocal segment of the word "fat", for example, after transformation is preceeded by a fricative and followed by a plosive,

although the individual phonemes no longer sound the same as the untransformed sounds. However the perception of phoneme order is sometimes destroyed in that two vowel-like phonemes merge into one diphthong*. The perceived uniqueness of phonemes is a rather complex aspect of cognition since there are no separations and the neighboring phonemes have a very strong influence. Thus, when the word "ball" is heard as "boy", one might say that the /ɔ/ and /l/ are merging into a diphthong /ɔI/. But, the diphthong, itself, might be considered to be two phonemes, /ɔ/ and /i/, if it were desirable to define it as such.

Questions about the use of spectral and nonspectral features as a means of identifying phonemes are again taken up in Chapter IV.

*See Section 4.22 for a discussion of the diphthong problem.

CHAPTER III

DESIGN OF EXPERIMENT

The experiment was divided into two sections: conversing and testing. During each session, subjects talked, or attempted to talk, to each other through the transformed medium for the first half-hour; they were then given a series of pre-taped tests lasting about a quarter hour. The conversation time provided the exposure to the transformed medium necessary for learning, while the testing series attempted to give an objective measure of learning and performance. Although the conversations were tape recorded and later analyzed for strategies, effectiveness, articulation, content and attitude, their subjective quality made comparison and evaluation difficult; so that, in effect, only the testing series provided an analytic measure of improvement.

The first practice session began with the following instructions:

> "Your goal is simply to learn to communicate with your partner using any approach you find practical. Eventually you should be able to communicate any new idea as if there were no spectral transformation present. The only restrictions are that you are not allowed to use a code which circumvents the transform, e.g. Morse code."

These instructions were repeated whenever subjects asked the author what they should be doing. In this way, subjects were allowed complete freedom to develop, test, and modify their own strategies of learning and communicating. Learning theories are too incomplete and contradictory to be used to indicate an appropriate teaching technique which could be applied to transformed speech. See Section 6.2.

A pilot study was performed prior to the main experiment in order to determine the frequency and length of the sessions necessary for attaining comprehension of transformed speech. At the end of five half-hour conversation sessions given on consecutive days, a pair of twins was able to converse fluently. Based on this pilot study, it was

decided that three sessions per week for seven consecutive weeks would be sufficient. The experience with the twins proved to be very misleading since they were very familiar with each other's speech habits and personalities. For this and possibly other reasons they proved to be the best, rather than a typical, pair of subjects. The questions of frequency and length of the conversation sessions are taken up again in Section 6.3.

3.1 EXPERIMENTAL ENVIRONMENT

The experiment was conducted in a small room which contained two acoustic chambers and the electronics to implement the spectral transformation. Each of the two subjects sat in a chair with his head inside a large wooden box lined with acoustical absorbing material. The acoustic tile and a layer of fiber glass served to isloate the two subjects by absorbing the speaker's voice and the room noise. They listened to the transformed speech through a pair of headphones embedded in a pair of "ear defenders". These ear defenders are used by jet airplane mechanics to reduce background noise. The combination of the headphones and the acoustic chambers prevented the subjects from hearing each other except through the transform electronics. The transformed speech, which normally appeared in the headphones, further masked any untransformed speech which might have leaked through. Even without the masking effect, the subjects could not hear each other directly. Also, they were visually isolated so that all communication had to pass through the electronics.

A microphone inside each acoustic chamber converted the subjects' voices to electronic signals. These signals were then added together, amplified, spectrally transformed*, and used to drive both sets of headphones. The acoustic levels were monitored with a VU meter and were generally kept at about 70 dB spl. Although the subjects were told to speak quietly and to keep their mouths about 6" from the microphone, they had a tendency to shout when they were not being understood. Since increased speaking effort often results in a decrease in intelligibility

*See Appendix B for a technical discussion and description of the transform electronics.

(Fairbanks and Miron 1957; House, Williams, Hecker and Kryter 1965), they were asked to speak softly rather than decrease the volume in the headphones electronically.

Even though both sets of headphones were driven by the same signal, subjects heard their own voices untransformed since bone conduction transmits the acoustic energy in the mouth directly to the middle ear through the headbone. With 30 dB attenuation of air conduction, a speaker's own voice is only reduced by about 6 dB (Békésy 1960, p. 187). Furthermore, all sounds are not heard equally as loud through the headbone. The phoneme /u/, for example, is heard rather loud as a result of a high sound pressure in the mouth increasing vibration of the lower jaw.

Periodic maintenance checks were run on the equipment to insure good performance throughout the experiment. The electronics were tested for the stability of the center rotation frequency, distortion, and leakage of the untransformed signal through the system. Of these, the last was most critical since the presence of normal speech, even in small amounts, would have invalidated the results by allowing a subject to understand his partner directly using the untransformed portion. The microphones and headphones were measured for frequency-response, distortion, and correct levels; also, the tape recorders were aligned and calibrated. Only once did a check reveal a malfunction. Interestingly, the failure was detected immediately by a change in the quality of the transformed speech. A transistor had failed.

3.2 CHOOSING SUBJECTS

The unstructured character of this experiment, in contrast to many other psychological experiments, required subjects who would be active participants rather than passive responders. For this reason, it was important to find subjects who would be highly motivated and interested in the experiment as an experience since the lack of specific instruction, the inherent difficulty of perceiving transformed speech, and the resulting frustrations suggest that a personality with ingenuity

and perseverance would be most suitable. Perseverance was evaluated by the potential subjects' persistence in requesting to participate in the experiment.

In a university where many psychological experiments are run, there exists a class of students known as "professional subjects". Many of these participate in experiments because they feel it is an easy way to make money; others participate because they might be "trying to prove something" to themselves. The relationship between personality and linguistic performance, underestimated initially, has been shown to be considerable. Lambert (1963) reports that students who are motivated to learn a foreign language with an integrative orientation were more successful than those with an instrumental orientation. The relationship between language performance and personality is again discussed in Section 6.2.

Although no systematic screening based on personality or language tests were given, when a subject volunteered for this experiment, after having heard about it from bulletin boards or from friends, he was told that he should participate if he thought he would enjoy it and that other experiments were far easier.

If a potential subject was still interested, he was given the base-line 45 minute test of session 1. Some became very interested in the experiment and would have been willing to participate without pay, although everybody was paid the standard $1.60 per hour; others showed a lack of interest and were never heard from again. Six out of the initial twelve subjects who took the base-line test were selected, more on the basis of scheduling than performance. They were told to find a friend with whom they could work as a team. Overcoming the spectral transformation is difficult enough without having to worry about talking to a stranger. In addition to being less inhibited about making mistakes with a friend, the subjects were mor. familiar with the partner's personality, interests, and speech habits. These extra-linguistic factors can be very important, as for example, having beforehand knowledge about the content of a speech utterance.

The experimental series began with six pairs of subjects, ten M.I.T. undergraduates and one pair of "intelligent hippies", with an average age of 20. The members of a pair were either close friends or members of the same living group, and on the average, they had known each other for about a year. Men were used exclusively in the experiment for three reasons. First, most students at M.I.T. are male and it would have been more difficult to find female subjects. Second, male speakers have a lower fundamental pitch which is richer in harmonic content and has a more pronounced envelop modulation. The individual pitch periods for a high pitched woman's voice tend to merge with one another making pitch more difficult to perceive. Third, the voice used to make the testing material was male. Since each of the subjects practice conversing with only one partner it was desired to make the testing voice as similar as possible to the subjects'.

The experiment continued for seven weeks with three sessions per week. Unfortunate scheduling difficulties sometimes resulted in either two or four sessions in a given week. Also, the final examination period prevented four of the pairs from completing the last three weeks of the experiment.

3.3 TESTING SEQUENCE

After every half-hour conversation session, each subject was given a quarter-hour series of tests to determine whether the additional exposure to the transform medium had produced improved performance on various speech tasks. The testing series attempted to be both an objective measure of the subjects' performance, and an indicator of which variables were responsible for learning to understand transformed speech. Since subjects could not initially communicate with each other and could do so after several hours of practice, some change in perception must have occurred. The results of the tests are the only available measures of the change.

During the first and last sessions, the subjects did not converse with each other, rather they were given a comprehensive 45 minute testing series. The results of the first session test established a base-

line level of performance with naive subjects. Although the subjects had never heard or practiced with transformed speech, some learning may have occurred between the beginning and end of this test. The identical tests were given at the last session so that an objective comparison, independent of test material, could be made. An unfortunate difficulty made the results of the last session somewhat awkward to use; four of the six pairs terminated after only four weeks. Thus, for two pairs the last session test was given on session 20, while for the other four pairs it was given after session 13. In some cases, the data is presented up to session 13, for others, the results of the last session are indicated as session "last".

Many of the tests were repeated throughout the duration of the experiment to avoid the problem of controlling for difficulty of content. However, the subjects were not informed of this fact, they were not told the correct answers, and they did not know how well or how poorly they had scored on the tests. For this reason, it was felt that repetitions did not introduce a strong bias. Subjects might remember an answer, consciously or unconsciously, if they understood a test word or sentence; but, if they understood it during the first presentation there is no reason to suppose they would not have understood it later, after more practice, had it been presented for the first time.

The kinds of tests used during each session varied, as can be seen from Table 3.3a. Four or five tests were given during any particular session. The testing series of sessions 11 through 19 were identical to those of session 2 through 10; and session 20 was the same as session 1. Within the series from sessions 1 through 10, some tests were repeated, with delays of typically a week. See the discussions of the specific tests in Appendix A for more details.

The test material was spoken by the author and, therefore, contains a certain amount of variability. This could have been controlled, or at least measured, by analyzing the recorded material for variations in such things as speaking rate, vowel-to-consonant energy ratio, pitch contour, loudness, etc. However, since this experiment was an

Session No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	last
Group I, Identification and Discrimination																				
Language discrimination	×			×			×			×			×			×			×	×
ABX consonant discrimination	×			×			×			×			×			×			×	×
ABX vowel discrimination	×																			×
ABX vowel control		×																		
Vowel confusion matrix	×		×			×			×			×			×			×		×
Consonant confusion matrix	×	×			×			×			×			×			×			×
Unvoiced plosive tests	×		×			×			×			×			×			×		×
Matched vowel identification	×	×			×			×			×			×			×			×
Group II, Comprehension																				
Words, one syllable	×			×									×							×
Words, two syllable	×	×									×									×
Words, geographic names	×		×			×						×			×					×
Words, in context					×		×							×		×				
Words, with categories								×	×	×							×	×	×	
Sentences, unrelated	×	×	×	×							×	×	×							×
Sentences, in paragraph								×	×	×							×	×	×	
Sentences, introductory clause					×	×	×							×	×	×				

TABLE 3.3a TESTING SEQUENCE FOR EACH SESSION.

exploratory study, it was felt that the large amount of work required to do this for all tests would not have been commensurate with the gains. Besides, only a small percentage of the features which contribute to natural speech are currently known, and therefore, one could not say when two stimulus words were equivalent.

Another approach would have been to use computer generated speech. However, because all the variables are controlled and because only those variables which are presently understood are used as parameters, computer speech is unnatural and presumably does not contain enough "human qualities" to be understood after transformation. The conclusion of the experiment is that speech has a rich enough internal structure to allow for comprehension even though some of the cues are lost. Computer speech is not complex enough to contain extra factors.

As a side note, it might be mentioned that each subject practiced only with his partner and was therefore familiar with only his transformed voice. It might be that a very low-pitch voice is particularly difficult to understand or that some other characteristic is particularly favorable. Moreover, the stimulus tests were made with the author's voice which might have been very different from the voice that subjects practiced with. Because of the limited experience with only one person, subjects could not generalize the effects of the transformation to other speakers. The test results could thus reflect a measure of the similarity between the partner's and author's voice.

The tests* listed in Table 3.3a are divided into two groups: 1) tests which measure the ability to discriminate or identify some speech feature, and 2) tests which measure comprehension under various conditions.

3.31 LANGUAGE DISCRIMINATION TEST

This test was designed to determine if subjects could perceive a "characteristic structure" in a language without necessarily understanding

*See Appendix A for the content of the tests.

the content or knowing which language was being heard. The passage shown below was translated into ten foreign languages and read by native women speakers.

> "Mary's teacher took her class for a nature walk one sunshiny day last week. Every time the group came to a new plant, they would stop and examine it while the teacher explained its parts. She showed them how a bee gets its honey from flowers and how a bug had eaten part of the leaves from some plants. On a few plants, the flowers had fallen off, and seeds had begun to form."

The thirteen versions of the passage, ten foreign languages, Arabic, Chinese, Japanese, Russian, Hindi, Rumanian, French, German, Hebrew, Finnish, and three English, were presented to the subjects in a random order. Their instructions were

> "You will hear samples of the same passage spoken in many different languages, some of which are English. If you think the passage is English write 'E', if not write 'No'."

On the first and last sessions, the complete text, lasting a total of 7 minutes, was used, whereas on the other sessions only the first 20 seconds of each version was included, making a total of about 3.5 minutes. The length of a complete version varied from a minimum of 18 seconds to a maximum of 40 seconds depending on the speaker. The reason for the truncation was to allow for additional tests to be included within the allotted testing time of 15 minutes.

Although the original experimental design called for using women speakers in this test, it was later decided that it would have been better to use men. As was previously mentioned, the lower vocal pitch of male speakers produces a speech which is rich in harmonic content. The envelop periodicity, with high pitched women's voices, is less pronounced after spectral transformation.

This test attempted to measure the subjects' sensitivity to

overall sound pattern differences, which, in the case of languages, are related to rhythm, prosodic features (stress, intonation, etc.) and phonemic structure. This contrasts with the phonemic discrimination tests which measure sensitivity to isolated sound elements. Judging whether a language is English is based on overall impressions with no specific criteria.

Another test (not listed in Table 3.3a), given on the last session, presented subjects with 14 samples of the sentence "Joe took father's shoe bench out." spoken by each of the subjects. Their instructions were :

> "You will hear the sentence 'Joe took father's shoe bench out.' spoken by each of the 14 subjects involved in this experiment. Indicate below your own voice by the letter 'M', your partners voice with a 'P', and a strange voice by 'S'."

Like the language test, this measures the subjects sensitivity to diffuse differences, but, in this case, the differences are a function of the idiosyncrasies in the subject's speech pattern and not of the particular language.

3.32 VOWEL AND CONSONANT DISCRIMINATION TESTS

Testing the subjects' ability to discriminate between two different phonemes gives a measure of the loss in perceptual sensitivity resulting from the spectral transformation. One could, just as easily, use another set of stimuli other than phonemes for measuring sensitivity, but this set is also related, in theory, to the sounds in a language. Determining the ability, or lack of ability, to make discrimination judgments is analogous to diagnosing a hearing defect. In fact, the idea and content of these tests was taken from the speech diagnostic tests of Robbins (1948). These tests try to answer the question "Which sounds can no longer be discriminated?".

Subjects were presented with 50 three word triplets at a rate of one triplet every 5 seconds. Two of the three words were identical, and

the third contained one phoneme which differed. In the consonant test, a typical triplet was "girl, girl, curl". The instructions for these tests were

> "You will hear groups of three words each. Indicate which of the three words is different from the other two by writing a 1, 2, or 3."

Since there are 22 English vowel phonemes, including diphthongs, in the International Phonetic Alphabet, it would have required 231 triplets to completely test for all possible discrimination. Only a selected representative sample was used because of the limited testing time.

A third discrimination test, listed as an ABX Vowel Control test in Table 3.3a, was used to examine the effects of stimuli ordering in the consonant and vowel tests. This test contained the eight triplets which were most frequently not discriminated on the vowel test of the first session repeated six times with every possible ordering. For a given pair of words "A" and "B", the six triplets, AAB, ABA, BAA, BBA, BAB, and ABB, were generated. A statistical analysis of the results using the Friedman two-way analysis of variance (Siegel 1956, p. 166) showed that there was no significant learning during the test, that is, subjects had the same probability of getting a triplet right at the beginning as at the end. It also showed that the results were independent of the ordering within a triplet.

3.33 VOWEL AND CONSONANT CONFUSION MATRIX TESTS

It is well known that with a set of acoustic stimuli subjects can distinguish many more than they can identify. The major difference between discrimination and identification is that in the former the two stimuli are immediately available in short-term memory while in the latter the subject must remember a set of reference sounds. Thus, the task of identifying spectrally transformed phonemes is related to being able to remember the new sound patterns. The fact that the sounds of speech, in contrast to an abstract set of sounds, are retained and do not obey Miller's magic number "seven plus or minus two" has been explained on

the basis of overlearning. It is therefore interesting to see if the spectrally transformed phonemes can be remembered.

In the vowel matrix test, subjects were presented with a sequence of 22 words in the form /b V t/ at a rate of one every five seconds. The set of stimuli consisted of "beet", "boot", "boat", "bate", "bat", "but", "bit", "bet", "bought", "Bert", and "bite". The consonant matrix test consisted of a sequence of 32 syllables in the form /C a/ taken from the following set of consonants: /b/, /d/, /g/, /p/, /t/, /k/, /f/, θ/, /s/, /ʃ/, /v/, /ð/, /z/, /ʒ/, /m/, and /n/. This is the same set used in the analogous experiment of Miller and Nicely (1955), thus allowing their results to be compared those obtained with the transformation.

The instruction given the subjects were:

"You will hear a series of words chosen from the following group. For each word you hear, put a cross in the box of the word spoken."

The instructions specifically ask subjects to indicate the word spoken and not the word perceived. The intention was to avoid the problem of being able to willfully shift the perception of a word, as for example, being able to perceive the spoken word "boot" as either "boot" or "beet". (See Section 4.23) If the subjects did not understand this phenomenon they would merely indicate the perceived word which they would equate as the spoken word; but if they were aware of the fact that they could change the perception they would indicate the word they thought was spoken without having to resolve the ambiguity in perception.

The test results, in addition to giving a measure of the ability to identify phonemes, provide the data for creating confusion matrices. In the Miller and Nicely experiment, the matrices showed that the confusion errors followed a definite pattern and that a feature discription of the errors was a useful structure. In other words, the error pattern is a way of determining if there are any physical and perceptual correlates of a hypothesized "feature set". For example, if a voiced phoneme is never confused with its unvoiced counterpart, such as /f/ and /v/,

then it may be said that there is both a physical and perceptual manifestation of the voicing-nonvoicing feature which is unaffected by the transformation. The converse, however, is not necessarily true; the lack of invariance does not prove that there is no feature since the transformation might have destroyed it.

3.34 UNVOICED PLOSIVE TEST

In order to explore the effects of the transformation on consonants within the same class, i.e. consonants with the same mode of articulation but different place of articulation, a special test was given to examine the perceptual behavior of the unvoiced plosive stop consonants in the initial position with different vowel contexts. On the first and last sessions, subjects were given a sequence 60 monosyllabic words with either a /p/, /t/ or /k/ as the initial sound. On their test paper appeared three possible choices for each stimulus word, e.g. for the stimulus "pole" there appeared "pole", "toll" and "coal" as choices. Their instructions were

> "You will hear a series of words. For each word underline the word spoken from the group listed below."

The second phoneme was varied from the ten choices, /o/, /i/, /æ/, /ɔ/, /ɛ/, /u/, /I/, /e/, /ʌ/, and /ɒ/. The manifestation of the plosive consonant is essentially a perturbation on the neighboring vowel, so that the vowel context should be important. This test is identical to the consonant matrix test except the consonant set is only three and the effect of the vowel context is explored.

During the other sessions, a different version of the plosive test was given in place of the above since that one was too long and it neither measured the subjects' ability to discriminate phonemes nor determined if an acoustic context could enhance perception. The vowel context was limited to only four vowels two back and two front, /u/, /ɔ/, /ɛ/, and /I/. Part I contained a sequence of 24 triplets, like those used in the ABX discrimination tests, with each of the four vowels used twice. The instructions were:

"You will hear a series of three words each. Indicate which of the three is different from the other two by writing a 1, 2, or 3."

Part II contained a limited version of the plosive test given on the first and last sessions using only four vowels. There were 12 stimuli with the instructions:

"You will hear a series of single words. Underline one of three words from each group that corresponds to the word spoken."

Part III was identical to Part II except that the stimulus word was preceded by a carrier, or introductory, sentence. See the next Section for a discussion of carrier sentences. The instructions for this part were:

"You will hear a series of single words similar to the above group, except this time the word will be preceded by the sentence 'underline the word spoken'."

3.35 MATCHED VOWEL TEST

A special vowel test was given in an attempt to examine in more detail the learning of spectrally transformed vowels. A group of word pairs, with the property that one word of the pair was perceived by naive subjects as the second word when spoken in the transformed medium, was found. For example, inexperienced listeners perceive the word "we" as "you". As a measure of the quality of the matching, it was found that the average score for naive subjects was approximately 12% even though random guessing would have produced a 50% score. Using this kind of stimulus set maximizes the likelihood of choosing the wrong response if one has not yet "adapted" to the medium.

In Part I, subjects were given the following instructions:

"You will hear a series of words each of which appears as one part of the following pairs. Underline the word spoken."

The stimulus words, in Part II, were preceded by a carrier sentence. The instructions were

> "You will hear a series of words similar to the above group, except this time each word will be preceded by the sentence 'Underline the word spoken'."

This part of the test is somewhat similar to those used in the experiments with synthetic speech, where it was shown that the perception of a stimulus word could be changed by the formant range of the carrier sentence (Fourcin 1968; Ladefoged and Broadbent 1957). In the experiment described by Fourcin the word "Hello" when spoken by a man, woman, and child influenced the perception of /b/ and /d/ (forced choice); Ladefoged and Broadbent showed that the range of the first and second formant in the carrier sentence "Please say what word this is." determined if a stimulus word in the form /b V t/ was heard as "bit", "bat", "bet", "but". The carrier sentence in the match vowel test is effectively determining if the carrier sentence provides a reference context which can be used to enhance identification. Since, under some condition, perception seems to be under the listener's control, the context might provide a strong cue to resolve the ambiguity created by being able to willfully shift perception.

3.36 WORD TESTS

Recognizing a word is the next level of complexity above identifying phonemes since it requires, or allows, the use of an internal dictionary. Word perception is not merely the sum of perceiving the individual phonemes for two reasons. First, phonemes do not occur in isolation and co-articulation effects modify phonemes considerably. Second, cues, such the number of syllables, accent, and context, limit the number of meaningful words which would be acceptable. These factors influence perception so that ambiguities arising from erroneous phonetic cues can be successfully resolved. With words, the percept is no longer completely dependent on the acoustic stimulus.

For the purpose of testing, words may be classified on the basis of their length since this is a cue which is clearly unaffected by the transformation. The words in the one- and two-syllable tests, taken from the Harvard PB lists (Egan 1948), had the same number of syllables (length) and no extra context cues. In the absence of context, probability of occurrence acts in much the same way as context because, according to Ziph's law (1949), the probability is generally a function of the number of syllables.

The two-syllable words were all two-word combinations, such as "houseboat", "blackboard", and "firefly". These "spondees", having homogeneous audibility, are all characterized by reaching the threshold of hearing within a very narrow range of intensity and have often been used to determine the threshold of hearing. Moreover, neither syllable is heavily accented. The one-syllable word lists did not contain any rare or unfamiliar words and, if conditions were chosen so that the average articulation score was 50%, very few of the words would be extremely difficult or extrememly easy to perceive. Also, the distribution of phonemes approximated normal English. The disadvantages of using these lists are discussed in Section 5.1. Each of the tests contained 20 stimulus words, presented at a rate of one every 5 seconds, and the instructions:

"You will hear (one)(two) -syllable words. Write each of them."

The stimuli in the geographic names test, in contrast to the word tests, contain cues based on the number of syllables, accent, and categorization. The names are the more common cities, states and countries, which represent a large but limited set. The results of this test when compared to those of the word tests demonstrate the importance of the extra cues. This test contained 20 names, present at a rate of one every 10 seconds, and the instructions:

"You will hear the names of cities, states and countries. Write each of them."

The words-with-categories test, like the geographic names test, attempted to measure directly the effect of knowing the context. Four sets of 10 words with common categories, consisting of household furniture, vegetables, colors, and parts of the body, were selected. On session eight 20 words composed of a mixture of the four categories were given in test which had the instructions:

> "You will hear a series of words. Write each of them."

The tests on sessions nine and ten, however, contained the separate groups with the instructions:

> "You will hear the names of (various parts of the body)(household furniture)(vegetables)(colors). Write each of them."

This allows the results to be compared under the two conditions: knowing the category to which the word belonged and not knowing the category. In theory, these tests served the same function as the geographic names test: namely, assessing the importance of non-acoustic cues on perception under conditions of spectral transformation.

The words-in-context test, given on session seven, was somewhat similar to the above except that the context was a sentence. Subjects heard a complete spectrally transformed sentence and which was printed on their paper with one major content word, generally the last, omitted. Their instructions were:

> "You will hear sentences as shown below. Fill in the missing word."

The sentences were taken from the specially generated set which has been used on articulation studies and might appear to be somewhat obscure, as for example, "A true saint is lean, but quite human.". The difficulty in using this kind of test is that the relationship between comprehension and the sentence content (and structure) cannot be evaluated.

Another words-in-context test, given on session five, explored the effect of an acoustic carrier sentence on comprehension. The stimulus words were the same as those used in the two-syllable tests and the instructions were:

> "You will hear two-syllable words preceded by the sentence 'Write down the word spoken.' Write down the word following the sentence."

This test served the identical purpose as the other carrier-sentence tests, mentioned in Sections 3.34 and 3.35.

One of the major difficulties with these tests was that the limited time allowed for word comprehension testing required that only one test be given on each session. However , a test which was a good measure of performance should have been given relatively often in order to gain useable estimates of learning, but a test whose results were conclusive and did not change with additional practice should have only been given a few times, preferably near the beginning and end of the experiment. Moreover, at any given time during the experiment, a test should have been neither so easy that all subjects performed very well, nor so difficult that no subject accomplished the task. Unfortunately, the experience with the pilot study created a false optimism and the appropriate difficulty was not always obtained. This was especially true with the sentence tests, as discussed in the next section.

To explore the phenomenon of learning spectrally transformed speech requires that the tests measure, or be a function, of those perceptual parameters which are changing. In one sense, this requires that one already know the answer in order to find the answer. This paradox was overcome by having a number of different kinds of tests which measure as many variables as possible. If the experiment were to be repeated the word tests would be modified to include others and omit some of those given.

3.37 SENTENCE TESTS

Comprehension in normal speech communications is more than the transmission of words; rather, it is the transmission of the ideas symbolized by the words and structure of the sentence. Because the interrelations between the words in a sentence are determined by the allowable syntactic structure and the semantic constraints, and because these aspects of communications are known by the listener, as well as the speaker, of a language, it is not unreasonable to expect that the cognitive processing of speech incorporates this knowledge to influence perception. Thus, any realistic hypothesis of speech perception must consider the semantics and syntax to be a set of conditions imposed on the perceptual response to the acoustic wave.

Using sentences to measure comprehension poses a difficult problem, namely, trying to use a stimulus which is not understood to measure a phenomenon which is not understood. For example, in making up a test for measuring comprehension under some condition, say spectral transformation, it is desirable to find a set of sentences which would be equally easy to perceive, yet, one cannot equate two sentences on the basis of difficulty unless one knows how the sentences are processed. Sentences, however, have the advantage of being a stimulus which is probably much more related to "essence" of communication than are isolated words.

There were three kinds of tests used in the experiment. The first set consisted of ten unrelated sentences with an average of five content words and three function words. The instructions were:

> "You will hear sentences repeated twice. Write down as much as you understand."

The sentences were repeated twice so that on the first listening the subjects would have time to absorb the stimulus as a unit and on the second listening they would write it word for word. Also, words perceived at the end of the sentence enhanced perception of the initial part of the sentence on the second presentation.

The original experimental design called for this type of test to be given throughout the experiment, but the performance level after the first week of the main experiment was so low that it was decided to use other kinds of tests. The design error was the result of an unrealistic expectation based on the experience with the twins in the pilot study. The purpose of the pilot study was to obtain an indication of the relative rate of learning and the relative difficult of the tests which should be used. Once the miscalculation became apparent, a second type of sentence test was introduced on session five.

In the second series of tests, the sentences in each session all contained the same initial clause. It was hoped that knowing the first part of a sentence would allow the subjects to "drift into the main clause". The instructions were:

> "You will hear sentences repeated twice. Write down as much of the second part as you understand. Each sentence begins with the phrase ('Because the weather was warm,')('During the month of May,')('Because school was closed,')."

The third series of tests was similar to the first series, except that the sentences formed a small story with each sentence relating to the previous one. This paragraph, in one sense, is a stimulus which is one level of complexity above single sentences. In a sentence, each word is embedded in the context of the neighboring words; in the paragraph, each sentence itself is in a context created by the theme of the sentences. Therefore, this test is an approximation to conversation. Subjects were given the instructions:

> "The following ten sentences are taken from a conversation between (two students on a Friday afternoon)(an automobile salesman and a prospective customer)(a young couple and a travel agent). The sentences will be spoken twice, the first time just listen to the conversation. The second time you will be given enough time to write each sentence."

On the first presentation the sentences were read without any pauses so that the subjects could listen to the general content, thereby acquiring the context or theme of the paragraph. The subjects were given enough time to write each sentence on the second presentation.

If the experiment were to be repeated, the three kinds of sentence tests would be interwoven instead of being given in groups. That is, the unrelated sentence tests should have been given on sessions one, four, seven, and ten, rather than on sessions one through four.

CHAPTER IV

PERCEPTION OF TRANSFORMED SPEECH

Phonemes, as linguistic building blocks, are symbolic elements which can be used to generate or transcribe speech. In much the same fashion, letters are elements in the alphabet set used to represent the visual analogue of speech in printed text. To a linguist a phoneme is a set of sounds, allophones, which "show characteristic patterns of distribution in the dialect or language under consideration" (Gleason 1955, p. 261). The concept of a phoneme as a set of sounds permits the initial, medial and terminal sounds associated with the graphic symbol "p" to be called the phoneme /p/ even though the sounds in the three positions are neither the same nor interchangeable. The definition of a phoneme is really a matter of pragmatism based on the notion that only by substituting a given phoneme with another phoneme can the meaning of a word be changed.

As one of the first steps towards understanding the perception of speech, the acoustic, perceptual and articulatory properties of phonemes have been extensively investigated during the last decade. Implicit in this approach is the assumption that the phoneme and its phonetic cues have a perceptual reality in normal conversational speech. It cannot be denied that the phoneme may well have a perceptual reality when used as an isolated stimulus in a context-free environment, but its role in the cognitive processing of normal speech has yet to be determined. The results described in this chapter are directed towards understanding the perceptual behavior of phonemes under the conditions of spectral transformation; however, the purpose of establishing, in great detail, the nature of this behavior is tc be able to show in Chapters V and VI that the phoneme performance is not the basis for learning to understand spectrally inverted conversational speech.

4.1 PHONETIC CUES AND DISTINCTIVE FEATURES

The process of characterizing phonemes is one of assigning

attributes to the physical manifestation of perceptual elements. As an ultimate goal, it would be desirable to establish a working definition of a phoneme which would allow a given physical stimulus to be classified on the basis of perception. Such a definition, in all probability, would be based on the cognitive mechanisms which cause a given sound unit to be perceived as a particular phoneme, or at least, with phonemes in isolation. However, this kind of approach regresses to the necessary-and-sufficient notion abandoned in Section 1.1.

An alternative and more prevalent viewpoint is to use the theoretical linguistic framework of "distinctive features". A set of distinctive features are a group of attributes which specify each phoneme in a language. Every phoneme in the linguistic set either contains, or does not contain, a particular feature and is marked with a + or - accordingly. No two phonemes have the same features. Although it was originally attempted to describe each binary attribute in terms of articulation, perception, and the acoustic manifestation (Jakobson, Fant, and Halle 1965), the feature set was later modified to simplify the transformational phonology of English and "universal" language. As a result, the relationship between distinctive features and the non-linguistic domains was considerably weakened (Chomsky and Halle 1968, cht. 7).

The binary feature concept, when applied to perception, has often been criticized, but Halle (1957) justifies the use of a binary rather than continuous space on the basis of the intended purpose of the feature system, namely, linguistics. While the linguists have become more conservative in the ways in which they are willing to use distinctive features, the psychologists have been incorporating this notion, more and more, into the design and analysis of experiments. Perhaps the apparent paradox of using linguistic features in a non-linguistic situation is best exemplified by the following comment of Denes (Hirsh 1966, p.112). "That is a question I ask him almost every time I see him (Jakobson), and after listening to his reply I am satisfied that I understand how he reconciles an acoustic definition of an essentially linguistic event. But a few days later all my doubts are back again."

Experimenters have tried, with varying degree of success, to relate the Halle feature system to perceptual correlates, acoustic parameters, and universals of articulation. Usually one is forced to modify the original set in order to apply it in a particular experiment. In the discussions on transformed vowels and consonants the notion of features is used but only when it can be demonstrated that this way of classifying the data reflects an underlying perceptual reality.

4.2 VOWEL PERCEPTION

4.21 UNTRANSFORMED VOWELS

Vowels, a subset of the phonemic classification of linguistic elements, distinguish themselves from consonants by the way in which they are produced. In each case, vocal cord modulation excites a mouth cavity which has no constrictions greater than that produced with the high vowels /i/ and /u/. The acoustic properties of the vowels can easily be modeled by viewing the mouth as a filter whose resonances are varied by changing the cavities created by the tongue (Fant 1960, cht. 1). Because of the predictable relationship between the filtering characteristics of the mouth and the spectrum of the sound wave, the acoustic nature of vowels is fairly well understood. The acoustic variables are usually specified by the frequencies of the resonance peaks, or formants, and by the bandwidth or damping of each resonance. Since the formant frequencies and amplitudes can be continuously varied from one vowel to another, there are an infinite number of vowel sounds that could be produced. In any given language, however, there are a limited number of phonetically defined vowels, but the difficulty that phoneticians have had in defining a unique set of English vowels with rules to determine how a given sound should be categorized attests to the lack of clear distinctions between similar vowels (Gleason 1955, p. 318).

Peterson and Barney (1952) reported that listeners classify vowels in isolation by the formant frequencies, although there was a significant overlap between neighboring vowels. Identification of vowels in context shows a more complex behavior. Using synthesized

speech samples, Stevens (1966) showed that the perception of the vowel in the word /b V l/ is significantly different from the same vowel segment in isolation. In a similar experiment, Lindblom and Studdert-Kennedy (1967) demonstrated that the identity of a vowel sound in a semi-vowel context is determined not only by the formants in mid-syllable but also by the direction and rate of transition. A consonant context, in addition to affecting perception of synthetic speech vowels, changes the fundamental frequency, loudness and duration in natural speech (House and Fairbanks 1953).

The formant pattern required for the perception of a vowel is relative rather than absolute. Shifting the formants by a fixed ratio, as is the case with vowels spoken by adult males compared to those of children, does not significantly change perception (Stevens and House, 1966). Moreover, Ladefoged and Broadbent (1957) showed that the perception of a vowel in the context /b V t/ could be made to be /bit/, /bet/, /bat/, or /but/ by varying the formant range of the vowels in the introductory sentences preceeding the test word. Further evidence of the non-categorical nature of vowels was given by Stevens, Ohman, Studdert-Kennedy, and Liberman (1964) in an experiment which demonstrated that there was only a slight increase in discriminability at the phoneme boundaries, contrasting sharply with the analogous situation with consonants. Since the slight increase in discriminability was found to be independent of language, they reasoned that the boundaries were not linguistically determined. In support of these findings, Fry, Abramson, Eimas and Liberman (1962) found that listeners are able to discriminate differences which are far smaller than would be needed in order to categorize a vowel and can discriminate more sounds than they can identify. It has been shown that the auditory system is very sensitive to changes in either formant frequencies (Flanagan 1955) or formant amplitude (Flanagan 1957). Also, preliminary investigations indicate that speech stimuli (consonants) are perceived better by the right ear (Shankweiler and Studdert-Kennedy 1967), whereas, vowels and sonar signals having a melodic base are perceived better by the left ear (Kimura 1964; Bryden 1963). These finding indicate that vowels are easily discriminable, non-categorical and heavily dependent on the vowel and consonant context.

The previous discussions suggest that it would be very difficult to describe vowel perception in terms of a characteristic feature system. There is, however, one vowel feature, tense-lax, which is not directly related to the steady-state formant pattern, but is a function of the manner of articulation. According to Stevens, House and Paul (1966), there are three acoustic correlates of this feature: 1) the formant frequencies of the lax vowel are closer to those of the neutral vowel /ə/; 2) the duration of the lax vowel is shorter; and 3) the formants of the tense vowel remain at their stationary values for a longer time. The duration correlate is clearly spectrally independent, and the other two are only somewhat affected by the spectral transformation.

The acute-grave feature (now called back-nonback), which is characterized by dominant high or low frequency energy, was never considered spectrally independent but subjects learned to compensate for the effects of the transformation.

4.22 DISCRIMINABILITY OF TRANSFORMED VOWELS

The spectral coloring of the vowels is changed by the system, although they maintain their essentially vocalic nature. Initially, some transformed vowels appear to be similar to some existing phonemes, while others are perceived as being either a combination of two vowels or totally alien sounds. Regardless of which phoneme a transformed vowel appear to resemble, it is perceived as having an unnatural quality characteristic of spectral transformation, and it is never thought to be normal speech.

A vowel discrimination test was given on the first and last sessions in order to determine if the set of vowels remained unique and distinguishable after transformation. Since Flanagan (1955) has shown that the JND for formant frequency differs for the first and second formants and is a function of frequency, it might be possible, as a result of a decrease in sensitivity produced by reversing the formants, for two vowels to become indistinguishable after transformation. Determining the ability, or lack of ability, of subjects to make discrimination

judgments is analogous to diagnosing a hearing defect with speech sounds.

In this test, discussed in Section 3.32 subjects were presented with sequence of 50 word-triplets. Two of the three words were identical, and the third contained a different vowel with the same consonants as the other two words. Subjects indicated which of the three words was different from the other two. The fact that the average score was 90% on the first session and 95% on the last session is strong proof that the transformation did not introduce a discrimination problem. All the vowel pairs, with a few interesting exceptions, were unanimously judged as being different after transformation. One would have predicted this based on the discussions in Section 4.21.

The difficult vowel pairs and their scores for a similar ABX test given on session two are shown in Table 4.2a. Although these pairs were more difficult to judge correctly, all but the last has a score which is significantly above chance. The first five pairs all contain back vowels. Apparently back vowels with a dominant band of low frequency energy are more difficult to perceive. No explanation is offered, although one might speculate that this results from the finer frequency distinction required when the spectrum is rotated. For the vowels /u/ and /o/, the first formants are 250 and 360 Hz, and the second formant are 700 and 800 Hz, based on perception with synthesized speech (Delattre, Liberman, Cooper and Gerstman 1952). Following spectral transformation, the first formant frequencies become 2950 and 2840, and the second formant frequencies become 2500 and 2400 Hz. On a ratio scale the differences are much smaller after transformation.

	% correct	p<
to-tow	56	.001
fool-full	75	.001
broke-brook	58	.001
pull-pole	49	.01
shoe-show	48	.02
man-men	60	.001
well-will	30	--

Table 4.2a Word pairs which are difficult to discriminate. The percentage correct is based on average score for 12 subjects and "p" is the probability that the score was the result of chance (expected value is 33%).

The difficulty with the last two pairs arises from a lack of stress. "It is well known that a vowel that is insufficiently stressed, in some sense, reduces to a mid or high central 'neutral' vowel" (Chomsky and Halle 1968, p. 59). Because the position of the tongue for the lax vowels is very similar, the spectra of the vowel sounds are not very different. In addition, the duration of the vowel is short and the formants may not have enough time to reach their final value. As a result, perception is very difficult since the differences are small and one does not have much time in which to perceive whatever difference exists. Even in the absence of spectral transformation one can perceive the differences between the tense vowels easier than between their lax counterparts.

It is curious to note that three of the pairs have the phoneme /l/ for their final consonant. In combination with this phoneme, vowels merge to form a diphthong which is perceived as a new linguistic unit. Furthermore, if the vowel is unstressed as is the case with the pair "well"-"will", the /l/ dominates the vowel contribution to the diphthong. The combination of a lax vowel in the presence of an /l/ explains why the score on the last pair was equal to change. With the second and fourth pairs, the vowel was tense and the diphthongs created were preceptually different. The phonemes /l/ and /r/ are, from an acoustic view, sufficiently similar to the vowels to be included with them.

The author found, using himself as a subject, that with extensive practice on these pairs one could learn to hear a difference, albeit a small one. Also, when the test was repeated on the last session subjects only made half as many errors as they did on the first test. For this reason, it is felt that the vowels remain completely distinguishable once the listener knows how to hear the differences.

4.23 CONFUSIONS OF TRANSFORMED VOWELS

Being able to distinguish the different vowel sounds does not mean that one can either recognize or remember them; moreover, with an arbitrary set of stimuli, of which transformed vowels might be an example, subjects can distinguish an almost infinite variety of sounds, whereas they can only identify about seven correctly (Miller 1956a). In the vowel recognition test, discussed in detail in Section 3.33, subjects were presented with a sequence of words in the form /b V t/ and were requested to judge which of the 11 possible choices had been spoken. From this data, a matrix showing the confusions between the vowel spoken and the vowel perceived was generated. By grouping the data in the large matrix into smaller matrices based on feature class, one can check for the existence of acoustic and perceptual correlates of transformed speech. Various reduced matrices for the data from session one are shown in Table 4.2b.

The distinctive-feature tense-lax, as shown in part (A) of the table, exists and is spectrally independent. Of all the features, this one has been highly defined in terms of the production, perception and acoustic domains. A tense phoneme, by definition, is executed with a "deliberate, accurate maximally distinct gesture that involves considerable muscular effort; nontense sounds are produced rapidly and somewhat indistinctly" (Chomsky and Halle 1968, p. 324). A tense vowel is longer in duration and often more intense than its lax counterpart. In the recognition test, the tense vowels had an average duration of 215 milliseconds, compared to 175 for the lax; the relative average intensity was 5.5 for tense and 4.5 for lax. The increased subglottal pressure in the production of tense vowels is directly related to the duration (Jakobson 1962) and often to the pitch (Ladefoged and McKinney 1963). Since duration and intensity are spectrally independent, it is not surprising that subjects perceive the difference between those vowels which are tense and those which are lax. The average number of errors in this feature, as shown on Graph 4.2b, decreased from 22% on the first session

		Perceived	
		Tense	Lax
Spoken	Tense	116	24
	Lax	37	82

(A)

		Perceived	
		Front	Back
Spoken	Front	1	50
	Back	51	14

(B)

		Perceived	
		Front	Central
Spoken	Front	9	27
	Central	26	21

(C)

		Perceived		
		High	Medium	Low
Spoken	High	38	27	5
	Medium	25	42	9
	Low	11	23	22

(D)

TABLE 4.2b REDUCED VOWEL CONFUSION MATRICES FOR SESSION ONE.

(A) Tense-lax.

Tense = beet, boot, boat, bate

Lax = bat, bet, but, bit

(B) Front-back with tense feature.

Front = beet, bate

Back = boot, boat

(C) Front-central with lax feature.

Front = bit, bet

Central = but, bat

(D) High-medium-low, relative position of tongue.

High = boot, beet

Medium = bate, but, boat

Low = bat, bawt

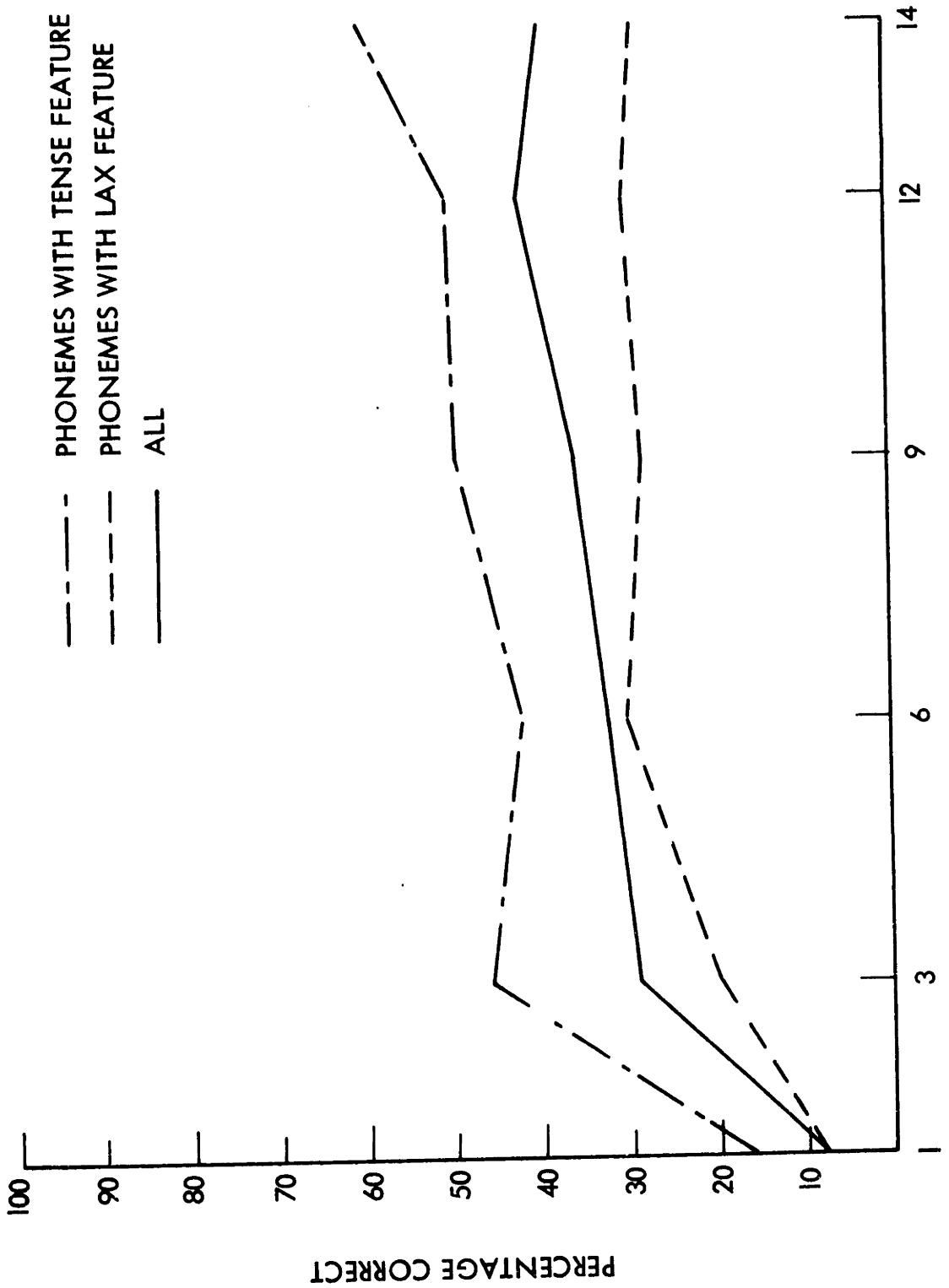
to 5% on the last session.* Thus, even on the first session, where the average number of absolutely correct answers was equal to chance, the tense-lax distinction was made.

In addition to a greater duration, a tense vowel is produced with a greater deviation from the neutral position of the vocal tract; and consequently the spectrum of the tense vowel shows a greater difference from the spectrum of the neutral vowel /ə/. Furthermore, "tense vowels have the duration needed for the production of the most clear-cut, optimal vowels, and in relation to them the lax vowels appear as quantitatively and qualitatively reduced, obscured, and deflected from their tense counterpart toward the neutral formant pattern" (Jakobson 1962, p. 552). This manifests itself by the fact that the tense vowels are consistently easier to identify correctly than the lax vowels, as shown in Graph 4.2a. The average score for correct identification of a lax vowel is asymptotic at 30% and reaches that value by session 6, whereas the average score for the tense vowels continues to increase with a final value of 60% on the last session.

Miller (1956b) also showed that the tense-lax feature resists distortion. In his experiment, he found that there were essentially no confusion between tense and lax vowels when listeners were presented with word in form /h V d/ which had been low-pass filtered to 670 Hz.

When the confusion matrix data from session one is reduced to the feature back-nonback (originally grave-acute) it reveals that this feature is spectrally distinct although not spectrally independent. Vowels with the back feature are produced by retracting the body of the tongue from the neutral position (Chomsky and Halle 1968, p. 304); and Delattre (1951) has shown that this corresponds to a lowering of the second formant. Clearly this feature is dependent on the shape of the spectrum of the vowel.

*Many of the calculations for the vowels omit subject pairs 1 and 2, since they never learned to perceive the vowels correctly. Including their results masks the point being made. The tense-lax errors become 35% and 12% rather than the 22% and 5%. This is discussed in Section 4.25.



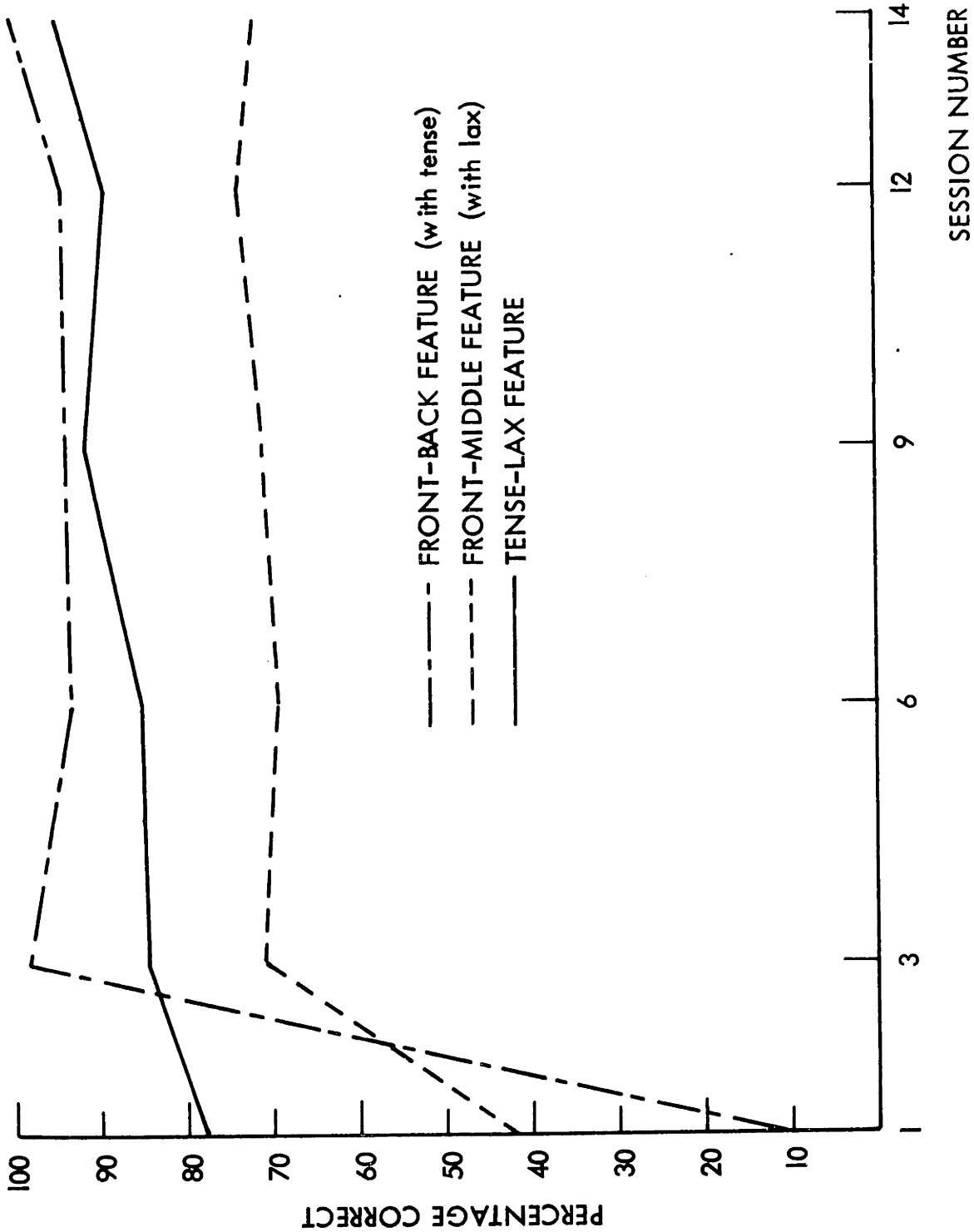
GRAPH 4.2a ERRORS IN REDUCED VOWEL MATRICES (without pairs 1 and 2).

On session one, the two front vowels in "beet" and "bate" were perceived as the back vowel in "boat"; the two back vowels in "boot" and "boat" were perceived as the front vowel in "beet". The judgments of back-nonback with vowels which also contained the tense feature were almost unanimous as can be seen in part (B) of Table 4.2b. This phenomenon is easily explained by the effect of the transformation on the more important second formant. For the vowel /u/ it is typically 800 Hz, and for the vowel /i/ it is 2200 Hz (Peterson and Barney 1952). The front and back vowels are acoustically reversed by the spectral rotation, and therefore perception is interchanged. The average correct score for this feature on session one was a mere 13%.

The most astonishing phenomenon was observed when the test was given a second time on session three. After only an hour of practice using the transformed medium, subjects compensated for the effect of the transform on the back-nonback feature and were able to perceive it correctly. The score on that session and all following sessions approached 100%.* This means that the perception of this feature is sufficiently plastic to allow for a complete reversal. No other sensory transformation has demonstrated such a rapid and complete compensation for pattern distortion (Harris 1965).

The compensation for the transformation immediately raises the question of what cues the subjects are using to re-invert the perceived spectrum. Firstly, subjects were told in the instruction to write down the word they thought was spoken. It is, therefore, not clear if the subjects actually perceived "boat" when "beet" was spoken, or if they knew, consciously or unconsciously, that, because of the alien nature of the sound, it must have been transformed and therefore the reverse of what it appeared to be. The author found that he could willfully perceive either "boat" or "beet" when the latter was spoken. This is discussed again in Section 4.24. Secondly, the sequence of test words provided a kind of context against which the subjects can reference a

*Subject pairs 1 and 2 are omitted since they never corrected for the transformation.



GRAPH 4.2b ERRORS IN REDUCED VOWEL MATRICES (without pairs 1 and 2).

test word. Ladefoged and Broadbent (1957) showed that the perception of /b V t/ test words could be changed simply by changing the range of the formants in the introductory carrier sentence. Stevens and House (1966) suggest that a vowel sound is referenced against the triangle formed by the three extreme vowels /u/, /a/ and /i/. Even though the explanation for the phenomenon is not clear, one can say that the vowel system is flexible and adapts to spectral transformation.

The analogous situation also occurs with the lax vowels, except that it is not as pronounced. Part (C) of Table 4.2b shows that there is a tendency for the lax front vowels to be perceived as central, but the central vowels are perceived equally as front and central. The perception of the back-nonback feature shows a similar reversal only much diminished. Graph 4.2b shows that the average score improves from 42% on session one to 71% on session three. The explanation is most likely based on the fact that the lax vowels are not very distinct since their spectra are very similar. This was also illustrated by the difficulty in distinguishing "men" and "man" on the previously mentioned discrimination test. Furthermore, since the lax vowels are short in duration, the formants never actual reach their final value. With normal speech, perception of the lax vowels results from extrapolating the final formant values by the rate and direction of the formants during the transitions to and from the neighboring phonemes. As will later be shown with consonants in Section 4.33, the ability to extrapolate the transition is severely impaired by the transformation.

Reducing the confusion matrix data with other features showed no significant perceptual correlates. One of these, degree of closure of the mouth cavity, is shown in part (D) of Table 4.2b. It does show, however, that the spectral transformation produces a shift in the direction of the high vowels. The major difference between vowels made with the tongue near the palate and those made with the tongue below the neutral position is simply the location of the first formant (Delattre 1951). For the high vowels it is as low as 160 Hz and for the low vowels it is about 600 Hz. It is hard to see how spectral transformation, which shifts these formants to 3040 and 2600 Hz, respectively, could

tend to make them be perceived as vowels with a low first formant. The explanation might lie with the behavior of the third formant, although this has not been investigated.

4.24 VOWEL INSTABILITY AND PHONETIC DICTIONARY

During the exploratory investigation of the transformed medium, the author attempted to create a phonetic dictionary showing how each phoneme was changed by the system. Even when a transformed sound was perceived as being similar to an existing phoneme, the perception could be shifted to the untransformed sound. The word "we", for example, could be perceived as its transformed equivalent "you", or it could be perceived as the word "we". Other vowel phonemes were often perceived as combinations of other phonemes. Because the perception of transformed speech appears to be under the listener's control, and because the phonetic alphabet contains only a limited number of entries without any rules for classifying unnatural sound, an exact phonetic mapping is very difficult or impossible to create. For some phonemes an equivalent representation was found, as shown in Tables 4.2c and 4.2d.

The plasticity in the perception of linguistic elements was also illustrated by Warren (1961) in an experiment in which listeners were presented with a continuously repeating test word. The word "tree", for example, during the course of the repetitions was heard as "tress", "press", "terez", "prez", "stress", and "teress". The semantic unity of the sound pattern disintegrated and allowed for new words to appear even though this required suppression, substitution, or addition of a new phoneme. Although this experiment is not directly related to the phenomenon observed with transformed speech, it does confirm the assumption that the auditory system is somewhat unstable in the perception of a linguistic sound pattern which contains no semantic information.

The switching perception of linguistic elements has several visual analogies. The "Necker cube" and the "Peter-Paul Goblet" are two figures which are perceived in two possible ways (Neisser 1967, pp. 90 and 144). In the former the figure changes perspective, in the latter

it changes in identity. Ambiguities in the perception of context-free stimuli exist in both the visual and auditory sensory modes.

One might assume that if one could pronounce a phoneme so that its spectrum was the reverse of what it should be then the system would rotate the spectrum so that the sound would appear normal. Subjects reasoned, incorrectly, that they could learn a phonetic transformation by imitating the sounds they heard when normal English was spectrally transformed. A subject would say "we" and hear "you" in his headphones; he then would say "you" whenever he wanted to communicate the word "we" to his partner. Although some subjects spent several hours trying to master this technique, they were forced to abandon it due to a lack of successful communication. The difficulty in using a phonetic transformation is explained by the lack of a one-to-one correspondence between a phoneme and its acoustic manifestation, and by the difficulty in mimicking a sound.

Richey (1968), in order to demonstrate spectral rotation as a privacy device for telephone conversations, created a phonetic dictionary which he used to speak in manner such that the resulting sound would appear normal after being spectrally transformed. After much practice, he had mastered this technique to the point where he could talk reasonably fluently. Unfortunately, his dictionary is no longer available and only the transcriptions of the nursery rhyme "Mary had a little lamb" and counting "one, two, three,.....ten" remain (Richey 1936). The dictionary of vowels, shown in Table 4.2c, was extracted from those transcriptions.

Although some phonemes occur in matched pairs, as illustrated, for example, by the fact that /i/ is perceived when /u/ is spoken and /u/ is perceived whn /i/ is spoken, many others do not. The phonemes /ɔr/ is heard when /ɛr/ is produced, but /ɛr/ is heard when /ɑr/ rather than /ɔr/ is produced. In this case the mapping is unilateral. Almost all the lax vowels are not bilateral and the choice of mapping appears to be dependent on the phonetic context in which the vowel occurs. The great variety of lax vowel pairs makes one wonder if there exists an

English Phoneme	Spoken Phoneme	English Word	Spoken Word	
aI	ɒ I	nine	malm	
aI	ɑ ʊ	white	yout	
o ʊ	e I	go	gay	Diphthongs
e I	o ʊ	ate	ohp	
ɜr i	ɔI I	Mary	Noyl	
I	j	little	ylippy	
I	j I	lamb	ylond	Semi-vowels
I	I	fleece	fludes	
w	j	went	yump	
ɛ	ʌ	ten	pun	
ɪ	ɪ	little	ylippy	
ɪ	ʌ	six	sux	
æ	ʌ	and	umb	
ʌ	ɛ	won	yen	
ʌ	o ʊ	a	oh	Lax vowels
ɛ	æ	every	adjew	
ə	u	the	thoo	
æ	ɒ	as	oz	
ɛ r	ɑ r	where	yahrr	
ɔ r	ɛ r	for	fair	
u	i	to	pee	
i	u	three	thru	Tense vowels

TABLE 4.2c PHONETIC DICTIONARY FOR SPEAKING TRANSFORMED SPEECH.

algorithm which yields the correct choice, or whether the choice does not matter very much. It is not obvious that an incorrect choice of a lax vowel would be perceived if the speaker maintained the right articulation rate and stress pattern. Unlike the lax vowels, the tense phonemes and diphthongs show a more clearcut bilateral mapping. The three pairs, /i/↔/u/, /eI/↔/oU/, and /w/↔/j/ are bilateral.

Context dependence is also exemplified by the many phonemes or phoneme pairs which are perceived as one sound. One hears the phoneme /aI/ when either /ɑ/ or /aʊ/ is spoken following an /m/ or /j/ respectively. The most complex phoneme in the dictionary is the semi-vowel /l/. The /l/ sound is created after spectral transformation by pronouncing /j/, /l/, or /jl/; but pronouncing /ɔIl/ creates the sound /ɛri/ which is not related to any of the three sounds which were used to create the /l/. The implication of this phenomenon is that speech cannot be treated as a sequence of perceptually distinct sounds which are transformed by the system in a simple one-to-one correspondence. This is in agreement with the general findings of speech synthesis experiments which have shown that the perception and production of a phoneme is influenced by the neighboring phonemes (Haggard and Mattingly 1968). In particular, the phoneme /l/ seems to have as many as six allophones and is almost impossible to represent with a single set of parameters without including modifications for initial, medial and final position, as well as for vowel and consonant context (Haggard 1967). Thus, a dictionary of transformed speech would have to include at least six entries for the phoneme /l/.

A third vowel test, described in Section 3.35 was used to explore some of the effects previously mentioned. A specially created set of word pairs was selected with the characteristic that the transformed version of one word should be perceived as the other. Subjects were presented with one word of each pair spectrally transformed and requested to indicate which word was spoken. The quality of the match was demonstrated by the lower than chance average correct score of 12% on the first session. Of the 21 word pairs, 15 had only one or no correct responses out of a possible 12. A dictionary of perceptual mapping,

as opposed to production mapping in Table 4.2c, was derived from the word pairs and is shown in Table 4.2d.

The trends are very similar to those observed by Richey; back vowels are generally perceived as front vowels and vice versa, and the /y/ glide becomes a /w/ glide. Chomsky and Halle (1968, p. 183) observe most tense vowels are diphthongized or have off-glides, /w/ glides for back vowels and /y/ glides for front vowels. Several other phoneme pairs were bilateral. Some of the pairs are not in agreement with those of Richey's dictionary but this may be the result of a different context, or it could be simply that the mapping for production and perception is not the same. That is, approximating the perception of a transformed English phoneme with another English phoneme may not yield the same results as trying to produce a sound which, when transformed, appears to be the desired English phoneme.

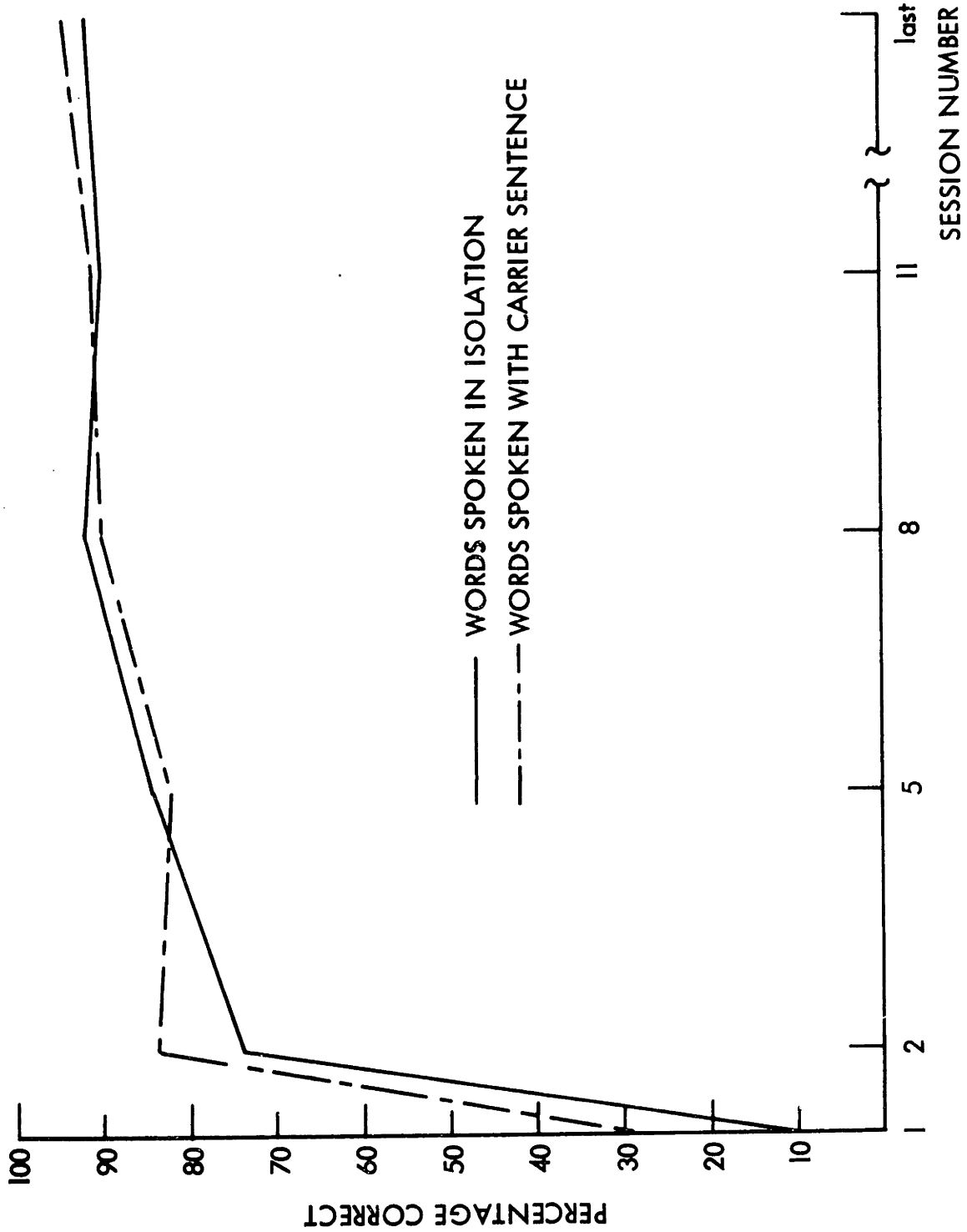
The phoneme pair /il/ and /ɔI/ appear to be bilateral in that one is perceived as the other; however, one should note that the test words in which these phonemes appear are not the same. This arises from the fact that "anneal" was perceived as "annoy", but "annoy" was not perceived as "anneal" even though "coy" was preceived as "keel". Again, it is found that /l/ is a complex sound which form diphthongs in combination with other vowels.

To generate a complete mapping table requires far more extensive experimentation. And furthermore, one must be careful that a subject does not begin to learn the transformation. Since it was found that an hour a practice significantly changed the subjects' judgments, only truely naive listeners can be used in evaluating a phonetic transformation dictionary.

The learning pattern in the correct perception of the matched word pairs is shown in Graph 4.2c. Notice that performance increased from a minimum of 10% on session one to a striking 74% on session two after only a half hour of exposure to the medium. This corresponds exactly to the phenomenon observed in the perception of the tense front-back

Stimulus Phoneme	Perceived Phoneme	Stimulus Word	Perceived Word	
ɪ l	ɔ ɪ	bill	boy	
i l	ɔ ɪ	anneal	annoy	
ɔ ɪ	i l	coy	keel	
r ɔ	j ɔ	raw	your	Semi-vowels
j ɔ	r ɔ	your ≠	raw	
j u	w i	you	we	
w i	j u	we	you	
a ɪ	o ʊ	dine	down	
ɑ ʊ	a ɪ	found	find	
e ɪ	o ʊ	ate	oat	Diphthongs
o ʊ	e ɪ	comb	came	
ɪ	ʊ	kid	could	
ʊ	ɪ	foot	fit	
ʌ	ɛ	pun	pen	
u	ɪ	loop	leap	Vowels
hir	h ɔ	here	whore	
ɔ r	ɛ r	door	dare	
ɛ r	ɔ r	mare	more	

TABLE 4.2d MATCHED VOWEL TEST GIVEN ON SESSION ONE.



GRAPH 4.2c PERCEPTION OF MATCHED WORDS (without pairs 1 and 2).

learning curve. Further practice results in a nearly perfect score.

This test was given in two parts; in part I the words were presented in isolation, in part II they followed an introductory carrier sentence, "Underline the word spoken". Having an acoustic context to which the test word can be referenced improved the performance significantly, although not dramatically, on the first session. With the sentences the score was 27%, compared to 10% without the sentences. Ladefoged and Broadbent (1957) found that the carrier sentence could affect perception by acting as a reference for the vowel formants in the test word.

The sensitivity to acoustic context may be one of the strongest cues in learning to correct for the spectral transformation. Stevens and House (1966) speculate that formants of the extreme vowels /u/, /a/, and /i/ are used to form a two dimensional scale that is used to normalize other vowels. Using a normalizing approach with the vowel triangle, Gerstman (1968) found the vowels of men, women, and children were accurately identified.

4.3 CONSONANT PERCEPTION

4.31 UNTRANSFORMED CONSONANTS

Consonants, unlike vowels, are much more categorical and possess easily definable binary features. The distinguishing cues for the perception of vowels are primarily spectral differences resulting from changes in the filtering characteristics of the oral cavity. The excitation source is always the modulated air stream produced by the spontaneous vibration of the vocal cords. With consonants, however, the primary pharynx excitation can be spontaneous voicing, non-spontaneous voicing, or an unmodulated air stream; moreover, there can also be secondary modulation such as plosion, frication, or nasal resonance produced by the articulators within the oral cavity.

Dividing the consonants into groups with common excitation is a

natural approach since the source differences are clearly observable in the acoustic waveform, as well as in both the manner of articulation and the judgments of perception. Miller and Nicely (1955) in their classical experiment found that the confusions among 16 consonants resulting from bandlimiting or adding noise were not between phonemes which differed in the source feature, but mostly between those with the same source feature. Most distinctive feature systems have very similar source features, usually referred to as manner of articulation, voicing and nasality, but they contain a variety of spectral or place of articulation features (Wickelgren 1966). Research with the perception of synthesized speech also shows that speech segments which are categorized as having the same source feature have many acoustic parameters in common (Cooper, Delattre, Liberman, Borst, and Gerstman 1952; Harris 1958). The validity of dividing the phonemes into groups is confirmed by the fact that many of the binary source features are directly implemented on speech synthesizers (Rabiner 1968) and vocoders (Schroeder 1966).

The five groups of source features used in the discussions of transformed speech are: unvoiced plosives, voiced plosives, unvoiced fricatives, voiced fricatives, and nasals. Nasals are executed with an opening to the nasal cavity, and their spectrum is similar to that of vowels and semi-vowels, in that the pitch pulses are just filtered, but the filtering includes the characteristic nasal resonances. Fricatives result from a constriction in the oral cavity sufficient to create the turbulence responsible for the noise quality. The voiced fricatives also contain a periodic pitch component. The plosive phonemes are characterized by a total closing of the vocal tract followed by a sudden release of energy; their primary cues, in a simplistic sense, are the silent stop interval and the rapid increase in sound pressure. Although the voiced and unvoiced plosives are distinguished by the presence or absence of vocal cord vibration, the perceptual difference is much more dependent on the additional energy and aspiration which occurs with the unvoiced phonemes, as well as on the duration of preceding vowel.

Using synthesized speech samples, Liberman, Delattre, Gerstman and Cooper (1956) have shown that the tempo of the formant transitions

is the most significant cue in being able to distinguish classes of speech sounds. The initial phoneme of a syllable was perceived as a voiced plosive, a semi-vowel, or a diphthong depending on the duration of the formant transition. By increasing the transition time from 10 to 300 milliseconds in steps, while preserving the formant frequencies, listeners judgments changed from /gɛ / to /jɛ/ and finally to /iɛ/. In the initial position, increasing the onset delay between the first and second formants changes the perception of a voiced phoneme to the unvoiced counterpart (Liberman, Delattre, and Cooper 1958). The duration of a silent interval with intervocalic plosives shows the same behavior; with a shorter duration it is perceived as a /b/, with a longer duration it is perceived as a /p/ (Liberman, Harris, Eimas, Lisker and Bastian 1961). Equally, /slIt/ can be changed to /splIt/ by inserting a silent interval of the appropriate length (Bastian, Eimas and Liberman 1961). In each case, the cue is temporal and, therefore, independent of the spectral transformation. One would expect that subject would have no additional difficulty in making these judgments with transformed speech.

The phonemes within a given class are executed with the same source feature but differ primarily in their place of articulation. The important acoustic cues for many of the consonants occur while the articulators are rapidly approaching an orientation of constriction. The location of the constriction determines how the accompanying vowel is affected. The cues which distinguish the phonemes are the rate and direction of the formant transition to and from the vowel. The difficulty in describing the acoustic parameters of the consonants arises from the fact that the transitions are dependent on which vowels are present. For example, a rising second formant with the vowel /i/ is perceived as /d/, but with the vowel /u/ it is perceived as /b/ (Delattre, Liberman and Cooper 1955). This situation occurs with most consonants (Liberman, Delattre, Cooper and Gerstman 1954); even some of the fricatives, which were thought to have been distinguished by their spectral coloring, have the formant transitions as a major cue (Delattre, Liberman and Cooper 1964). Similarly, Lisker (1957) has shown that the perception of the semi-vowels in intervocalic position is significantly affected by changing the vowel context.

In contrast to the perception of vowels, which is continuous, the perception of consonants is very nearly categorical. Experiments with synthetic speech have shown that stimuli, varied along an acoustic continuum, were virtually indistinguishable except at phoneme boundaries (Liberman, Harris, Hoffman, and Griffith 1957). In other words, subjects could distinguish two speech-like stimuli only if they could identify them as being different linguistic elements. No such peaks in discrimination have been found when non-speech stimuli are used. Stevens and Halle (1957) and others have used this observation as the basis for the argument that there exists a speech mode of perception. After a long exposure to a language, presumably mostly during the acquisition stages, the listener develops an "acquired distinctiveness" which allows him to divide the incoming stimuli into groups or compartments with a unique linguistic label. The boundaries of the compartments would then be dependent on the particular native language. "It has been speculated that this learning process is assisted because the listener, who is, of course, also a generator of speech, has available to him the transformation between auditory patterns and instructions to the vocal mechanism that give rise to these patterns" (Liberman, Cooper, Harris, MacNeilage and Studdert-Kennedy 1967). The difficulty in making linguistic judgments in a newly acquired foreign language with different parameter boundaries for the linguistic compartments attests to the permanence and stability of the consonant categorization.

This has several implications for transformed speech. Since the spectral transformation changes the formant pattern by inverting the rising and falling transitions as well as shifting the location of the formant, one would predict, on the basis of the stability of consonant categorization, that listeners would never be able to learn to make correct identification of consonants which differed only in the place of articulation. Or at least, the task would be as difficult as becoming a fluent speaker of a foreign language.

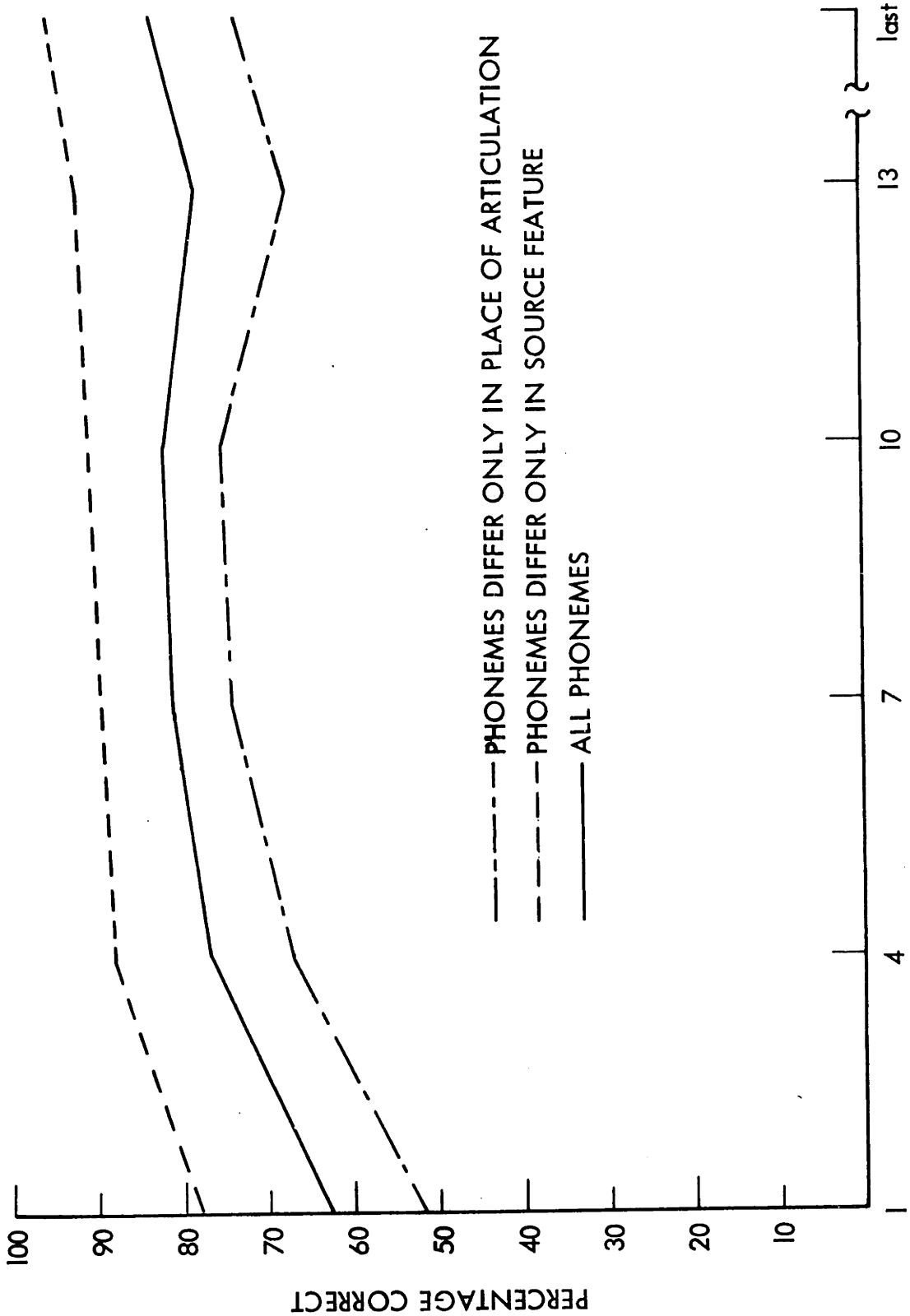
4.32 DISCRIMINABILITY OF TRANSFORMED CONSONANTS

Consonant differences are much more difficult to perceive than

vowel differences since consonants are low in energy content, rather short in duration, and often owe their perceptual existence to the perturbations of the neighboring vowels. For example, the intensity of /θ/ is about 27 dB less than /ɔ/ (Fletcher 1953, p. 86); the duration of the noise burst and formant transition for the plosive phonemes is about 50 milliseconds (Liberman, Delattre, Gerstman, and Cooper 1956) compared to about 200 milliseconds for a tense vowel (House and Fairbanks 1953). An ABX-type consonant discrimination test, described in Section 3.32, was used to measure the sensitivity to phoneme differences after spectral transformation.

Subjects were presented with groups of three words, one of which differed from the other two in that it contained a different consonant almost always in the initial position. They indicated which of the three they perceived as being different. As shown in Graph 4.3a, the results indicate that it is much harder to perceive the difference when the phonemes differ in the place of articulation rather than in the source feature. For example, "curl" and "girl" were always perceived as being different, whereas this was not true of "pan" and "tan" which were virtually indistinguishable. This phenomenon occurred for plosives, fricatives, nasals, and semi-vowels. The prediction of the Section 4.31 is borne out; the source feature of the consonants is more stable and more resistant to distortion since the acoustic cues are generally non-spectral.

The sensitivity to consonant cues increased rather significantly by the second time the test was given on session four and remained essentially constant thereafter. In contrast to the score for the source feature, which approaches 100%, the score for the place of articulation feature appears to have had an upper bound of 75%. Thus, if the test is thought of as a diagnostic measure of speech performance, one can say that spectral transformation introduced some speaking or hearing loss, depending on the point of view. Although the effective hearing loss did not approach a minimum value, at least after ten hours of conversational exposure, one can not predict the score for a subject who might have had over 1000 hours of exposure.



GRAPH 4.3a ABX CONSONANT DISCRIMINATION TEST.

The results of this test must be qualified for several reasons. The sample words used were read by the author and they were therefore not controlled for speaking rate, intensity, pitch, or other variations. Although an attempt was made to keep pronounciation constant, the natural differences which exist between the same word spoken twice would tend to hinder correct judgments. On the other hand, excessive accentuation of the dissimilar word would enhance perception of the different phoneme. It has been shown that the intelligibility of an utterance is a function of the consonant-to-vowel energy ratio (Williams, Hecker, Stevens, and Woods 1966, p. 23). And, increasing vocal effort decreases the intelligibility when the average intensity is held constant (Fairbanks and Miron 1957). The difficulty in controlling for pronounciation variations using a computer or spectrograph is simply that only some of the variables involved in speech perception are understood. Equally, there is no "standard" pronounciation for a phoneme or syllable.

The existence of a consonant is most clearly perceived when the sound is heard as a word with semantic content. Since subjects did not understand the stimulus words or even attempt to categorize the segments as linguistic elements, they were essentially judging the differences in non-speech samples. In essence then, the test measures the ability to perceive differences in the perturbations of the vowel formants, that being the manifestation of many of the consonants with the same source feature. In addition, the nature of the formant transitions is heavily dependent on the particular vowel following the consonant (Delattre, Liberman, and Cooper 1955); and differences between two consonants might be perceived with some vowels and not with others. The situation is further complicated by the fact that the perception of changing resonant frequencies, formant transitions being an example, is very much dependent on whether the listener is operating in a "speech mode" (Brady, House, and Stevens 1961). Perception is significantly enhanced if the elements are identified as speech sounds.

There are other cues besides the formant transitions which could be used by the subjects in distinguishing consonants. Their energy content differs by amounts greater than the JND for loudness differences,

although the relatively large intensity of the vowels tends to mask the consonants. In addition, the secondary characteristics of the vowels, such as pitch, duration and intensity, are affected differently by the different consonants (House and Fairbanks 1953).

As a final qualification, it should be mentioned that the 3000 Hz bandlimiting of the spectral transformation itself removes the high frequency cues in the speech signal. This is particularly noticeable with the unvoiced fricatives, which are often indistinguishable when the signal is merely bandlimited, without any spectral transformation (Heinz and Stevens 1961).

4.33 CONFUSIONS OF TRANSFORMED CONSONANTS

Correct identification of consonant phonemes requires that the listener preceive all the cues and features necessary to make a unique judgment. In contrast, discrimination of complex stimuli, such as consonants, can be performed if they differ in only one binary feature or if one feature has a different manifestation in the acoustic waveform. Identification can never be better than discrimination and the maximum correct score for the latter was only 83%. Nevertheless, examining the errors made in consonant identification can give some insight into which features resist destruction by the spectral transformation.

Subjects were given a sequence of syllables in the form /C a/ and requested to indicate which consonant was spoken. In order to compare the results with spectral transformation to those with additive noise and bandlimiting, the same 16 consonants, /p/, /t/, /k/, /b/, /d/, /g/, /f/, /θ/, /s/, /ʃ/, /v/, /ð/, /z/, /ʒ/, /m/, and /n/ of the Miller and Nicely experiment were used. See Section 3.33 for details.

The confusion data from the test given on session one were used to generate the reduced confusion matrices shown in Table 4.3a. Part (A) of the table is the confusion matrix for the combined primary and secondary source features, while parts (B) and (C) show the matrices for these features separately. The overall error rate for confusing a

(A)

Spoken \ Perceived	P_n	P_v	F_n	F_v	N
P_n	68	3	0	0	0
P_v	0	43	1	8	3
F_n	3	5	66	15	0
F_v	1	24	13	46	13
N	3	0	3	7	31

(B)

Spoken \ Perceived	N	V
N	137	23
V	21	175

(C)

Spoken \ Perceived	P	F	N
P	114	9	3
F	33	140	13
N	3	10	31

TABLE 4.3a REDUCED CONSONANT CONFUSION MATRICES FOR SESSION ONE.

(A) primary and secondary source features.

P_n = unvoiced plosive, P_v = voiced plosive,
F_n = unvoiced fricative, F_v = voiced fricative,
N = nasal

(B) primary source feature.

N = unvoiced, V = voiced

(C) secondary source feature.

P = plosive, F = fricative, N = nasal

phoneme with one having a different source feature is only 29%; yet the average error rate for identification is 77%. The source feature errors, when further separated into primary and secondary, show that confusions between the voicing and non-voicing feature is only 15%, and between plosive, fricative, and nasal is 20%. This clearly indicates that, without any previous exposure to spectral transformation, subjects perceived the acoustical manifestation of the source features, but did not perceive the place of articulation feature. This is in complete agreement with the trend found in the consonant discrimination test, and also confirms the conclusions of Miller and Nicely (1955). They found that the most stable feature with bandlimiting and/or additive noise was voicing-nonvoicing; the least stable was the place of articulation.

The same data can also be used to check for the existence of other distinctive features which have been postulated in various experiments and theories. Duration and place of articulation have been used by Miller and Nicely in their experiment of consonant confusions; openness of the vocal tract and place of articulation were used by Wickelgren (1966) in his experiment of confusions in short-term memory during recall; strident, grave, diffuse, and continuant have been used in the linguistically based distinctive-feature system (Jakobson, Fant, and Halle 1965). However, since most of the consonants have been shown to have no acoustically invariant cue (Liberman, Cooper, Shankweiler, and Studdert-Kennedy 1967), it would be surprising to find any consistent pattern in the confusions of spectrally transformed consonants.

It has, however, been found that the locus of the second formant does appear to be somewhat independent of the phonemic context. The /d/ seems to have an imaginary 1800 Hz starting second formant frequency, while the /b/ is somewhat lower and the /g/ somewhat higher (Delattre, Liberman, and Cooper 1955). This is equally true of the unvoiced plosives and nasals. Delattre, Liberman, and Cooper (1964) showed that the fricatives also possess an invariant locus for the second and third formant transitions. For /f/ and /v/ the second formant locus is 700 Hz, for /θ/ and /ð/ 1400 Hz, for /s/ and /z/ 1600 Hz, and for /ʃ/ and /ʒ/ 2000 Hz.

Since the locus is the formant frequency which would have been produced had any sound been generated at the time of constriction, the place of articulation feature is directly related to the locus value. Sounds which are made with a front constriction have a lower frequency locus than those made with a back constriction. If the perceptual system is generalized to the degree that the detection of the consonants is based on the extrapolation of the existing formant transition to find the locus, one could predict that the confusions of the consonants would be consistent and independent of the vowel context.

The confusions of a particular consonant with others having the same source feature appear to be almost random with no observable pattern. Also, the unvoiced plosive test described in Section 4.34 conclusively demonstrates that the place of articulation feature is almost impossible to perceive consistently and that it is very context dependent. When the data is reduced to the place-of-articulation matrix of Table 4.3b, one observes that there is a slight perceptual shift in the direction of front constriction. With this exception, consonants with the same source feature are virtually indistinguishable, either because the formant transitions are not longer perceived or because some other cue, such as loudness, dominates any judgment based on formants.

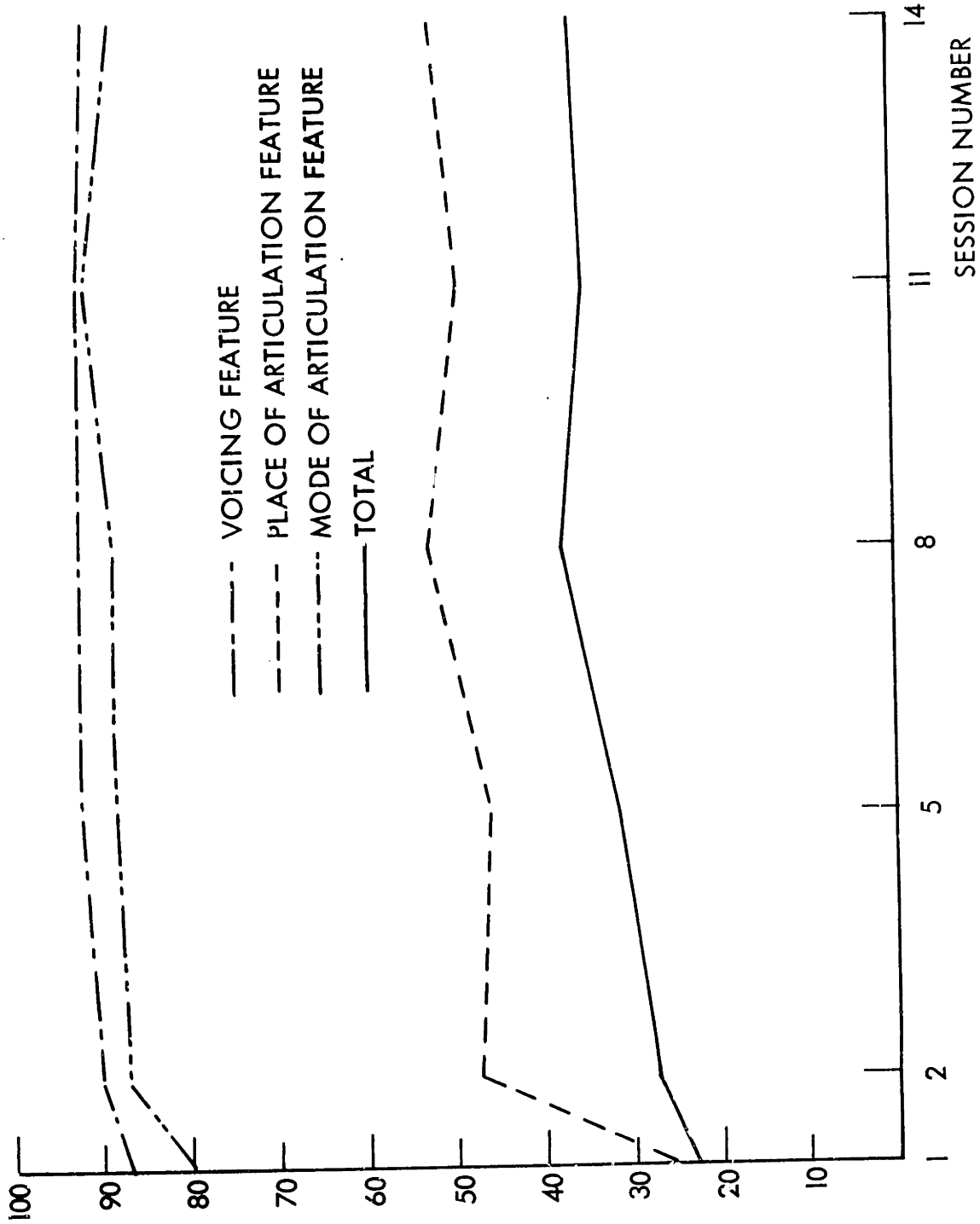
The ability of the subjects to correctly perceive the place of articulation feature, as shown by the learning curves in Graph 4.3b, improved quite significantly between sessions one and two, and improved slightly between sessions five and eight. The average correct score for this feature increased from 25% to 48% by session eight; thereafter it remained approximately constant. Although the form of the curve is very similar to that of the place of articulation for vowels in Graph 4.2b, in that both have the most significant improvement during the first few sessions, there is no reason to suppose that the same mechanisms of perception are involved. The consonants are far to complex to be able to state which cue was responsible for the change. The improvement might simply be the result of using the differential loudness of the consonants with the same source feature in making educated guesses.

		Perceived F	Perceived M	Perceived B
Spoken	F	28 (35)	53 (35)	21 (33)
Spoken	M	43 (37)	55 (35)	14 (38)
Spoken	B	41 (38)	51 (38)	44 (61)

TABLE 4.3b CONSONANT PLACE OF ARTICULATION CONFUSION MATRIX FOR SESSION ONE.

Front = /p/, /b/, /m/, /f/, /v/
Middle = /t/, /d/, /n/, /θ/, /ð/
Back = /k/, /g/, /s/, /z/, /ʃ/, /ʒ/

(..) = expected value assuming perfect score for source features and chance for place of articulation feature.



GRAPH 4.3b ERRORS IN REDUCED CONSONANT MATRICES.

In fact, the correlation studies in Section 4.6 strongly suggest basic differences between the perception of vowels and consonants. Nonetheless, it is curious that the most significant improvement occurs only for the first hour of exposure to the transformed medium.

The source features, having spectrally independent temporal cues, were clearly perceivable on the first session, and showed a slight improvement after a half hour of exposure. As can be seen in Graph 4.3b, errors in the primary source feature, voicing-nonvoicing, approached 8%, and the secondary source feature, plosive-fricative-nasal, approached 10%. The greatest source of errors for the latter features was the voiced fricatives were often perceived as nasals or voiced plosives.

Voiced fricatives are characterized by a periodic component and a noise component. The percentage of the energy that is noise is determined by the degree of constriction; with a very narrow constriction, the differential glottal pressure decreases, the voicing becomes non-spontaneous, and the sound appears to be similar to the unvoiced fricative. With a weak constriction, turbulence is reduced and there is a corresponding reduction in the noise component. Again, it is found that variations in pronounciation have a strong affect on perception. A similar effect can be found with the plosives. The phoneme pairs, for example, /b/ and /p/ are distinguished by the voicing feature, but the perceptual difference is much more dependent on the additional energy and aspiration that occurs with the /p/ (Halle, Hughes and Radley 1957).

Thus, the source features, as acoustically realized, are not truly binary and the perceptual distinctiveness is dependent on the speaker's articulation. In general, however, they are stable and unaffected by the spectral transformation.

In contrast to the learning curves for the features, the score for the absolute correct consonant identifications increased linearly from 23% on session one to 36% on session eight and remained constant

thereafter. Since this curve does not follow the feature curves, one could speculate that perception of the features does not lead directly to the perception of the phoneme. Perhaps, additional exposure to the medium is required to integrate the perception of the features, or even more heretically, perhaps the feature detection is not the basis of perception.

4.34 PERCEPTION OF UNVOICED PLOSIVE PHONEMES

In order to evaluate in detail the effects of the spectral transformation on consonant phonemes with a common source feature, a series of tests using unvoiced plosives was given. Since these consonants manifest themselves in the acoustic wave as a noise component and a formant transition, they are particularly appropriate choices. The testing series described in Section 3.34, attempted to examine the subjects' ability to discriminate and identify phonemes as a function of the vowel context, exposure to the transformed medium, and acoustic carrier sentences.

According to the synthetic speech experiments of Cooper, Delattre, Liberman, Borst, and Gerstman (1952), the acoustic cues for the perception of the unvoiced plosives are the frequency of the noise burst and the extent of the formant transition. The aspiration cue for the /t/ lies above 3000 Hz and is therefore removed by the band-limiting. The noise burst for the /k/ is approximately equal to the second formant for back vowels and above it for the front vowels. Any other frequencies for the noise burst contribute, to a greater or lesser extent, to the perception of the /p/. They also found that a rising second formant transition is always perceived as a /p/ independent of the vowel; but a falling transition is perceived as a /k/ for front vowels and as a /t/ for back vowels. Generally, the lack of a transition is perceived as a /t/ for front vowels and as a /k/ for back vowels.

A discrimination test was given to determine the subjects' sensitivity to the differential cues of the transformed plosives. The test, in an ABX format, contained a sequence of three-word groups; two of the

three words had the same plosive. Four different vowels, two front and two back, followed the plosive so that any effect from context would be eliminated. The average discrimination score increases monotonically from 42% on session three, which is only slightly above chance, to 56% on session twelve. This learning curve shows no indication of having reached a maximum limit, although there is insufficient data to predict its behavior for additional exposure to transformed speech.

Of the four vowels /u/, /ɔ/, /ɛ/, and /I/, the discrimination score with the last was consistently and significantly better than with any of the others, approaching a maximum value of 70%. It is difficult to explain why this should be so. Also, it was far easier to discriminate /k/ from /p/ than either of the two other possible combinations. Parenthetically, the locus of the second formant for the /k/ and /p/ is maximally distant.

An identification test consisting of a sequence of 60 words each with a different initial plosive and any of ten possible vowels was given on the first and last sessions. Subjects indicated which of three possible words, differing in the initial phoneme, they thought was spoken. The resulting confusion matrix shows a very complex pattern. From the summary matrices for front and back vowels as shown in Table 4.3c, one observes that the vowel can have a very pronounced effect on perception, especially for the /k/.

In trying to formulate a technique for reading the spectrograms of the plosives, Potter, Kopp and Kopp (1966, p. 103) observed that the /k/, when followed by a front vowel, has a high virtual second formant, a higher frequency noise burst, and tends to merge the fourth and third formants of the vowel; but when followed by a back vowel, it has low virtual second formant, and tends to merge the second and third formants of the vowel. The complete dependence on the vowel gives a basic explanation for the fact that it is correctly perceived 92% with back vowels and only about 15% for front vowels (on the last session). This

First Session

Followed by Back Vowel	Spoken	Perceived /p/	Perceived /t/	Perceived /k/
	/p/	44	20	10
	/t/	46	17	11
	/k/	25	11	39

Followed by Front Vowel	Spoken	Perceived /p/	Perceived /t/	Perceived /k/
	/p/	16	45	31
	/t/	38	18	35
	/k/	63	20	11

Last Session

Followed by Back Vowel	Spoken	Perceived /p/	Perceived /t/	Perceived /k/
	/p/	35	21	40
	/t/	46	20	30
	/k/	4	3	89

Followed by Front Vowel	Spoken	Perceived /p/	Perceived /t/	Perceived /k/
	/p/	33	43	43
	/t/	43	44	33
	/k/	74	24	22

TABLE 4.3c CONFUSION MATRICES FOR UNVOICED PLOSIVE PHONEMES.

effect is also present in the results of the first session but to a much smaller degree.

On the first session, perception of the /p/ is generally shifted to a /t/ and somewhat to a /k/ for front vowels; for back vowels it is perceived as itself, /p/. Similarly, the /t/ is perceived as both a /p/ and /k/ for front vowels and as a /p/ for back. The /k/ is perceived as a /p/ for front vowels and as itself /k/ and somewhat also as a /p/ for back vowels. The exact matrices for this data, shown in Table 4.3c, show that the average correct score is only 16%, significantly less than chance, for the front vowels; but it is 45% for the back.

Increased exposure to the transformed medium does not seem to have improved perception of the /p/ and /t/ with back vowels, although the /k/ showed a very marked improvement. On the other hand, the plosives with a front vowel showed a general improvement with the average correct score increasing from 16% to 28%, which is approaching chance.

A third test attempted to determine whether having an acoustic context, such as a carrier sentence, would aid in correct identification as it had in the matched vowels test. The sentence "Underline the word spoken" preceded each test word. The results from this test were conclusive: the carrier sentence produces no statistically significant change in perception.

Several conclusions may be drawn from the results presented in this section. Firstly, it is difficult to differentiate between two transformed plosives. Using the speech mode of perception model discussed in Section 4.31, which assumes that the differences between two stimuli must lie across the phoneme boundary if the stimuli are differentiated, one observes that the categorization boundaries for the plosives are shifted. Furthermore, the plosive stimuli in the synthetic speech experiments were generated by varying only one parameter, the second formant transition, and the perception judgments of this parameter were found to be categorical. One might speculate that

perception of the many cues in natural speech, such as noise burst frequency, pitch changes, second and third formant transitions, is also categorical; then, correct identification would be based on a reasonable agreement between the many cues. One must realize, however, that the spectral transformation shifts all category boundaries relative to each other. No concensus would ever be achieved between the judgments except, possibly, for some special combinations which happen to fit those of natural speech.

Secondly, the behavior of the transformed consonants shows that the perceptual mode does not extract an invariant cue from the stimuli since the subjects' judgments are rather inconsistent. Thirdly, the stability of categorization in the hypothesized speech mode of perception is born out, in that identification judgments do not significantly improve even though discrimination between differences does. The change in identification judgments after 10 hours of exposure might show that subjects are learning to use different types of cues rather than relearning the use of old features. Fourthly, sensitivity to the differences in plosives is very dependent on the vowel and on pronounciation. This would confirm the observation that consonants are very complex.

Fifthly, it would be fairly safe to state that, in general, perception of the place of articulation feature, at least for unvoiced plosives, is destroyed by the spectral transformation.

It is interesting to note that Richey's (1956) phonetic dictionary for speaking transformed speech confirms many of the observation made in the previous sections. All the phoneme pairs that map from normal speech to transformed speech preserve the source features, but there does not seem to be any observable algorithm for generating the place feature. The phoneme pairs shown in Table 4.3d are included as an illustration since there are far toofew to use in the creation of a consistent pattern.

English Phoneme		Spoken Phoneme	English Word	Spoken Word	
d		b	had	hob	
b	?	d	lamb	ylond	Voiced Plosives
g		g	go	gay	
t		p	little	ylippy	
t		t	white	yout	Unvoiced Plosives
m		n	Mary	Noyl	
n		n	won	yen	Nasals
n		m	nine	malm	
f		f	four	fair	
s		s	snow	smay	Unvoiced Fricatives
ʃ		θ	sure	theer	
v		v	seven	suvum	
v		dʒ	every	adjew	
z		z	whose	heez	Voiced Fricatives
ð		ð	that	thop	

TABLE 4.3d PHONETIC DICTIONARY FOR SPEAKING TRANSFORMED SPEECH.

4.4 PROSODIC FEATURES

The ease with which subjects learn to communicate in transformed speech can be understood by observing that the many cues which contribute to the perception of the prosodic features remain unaffected by the system. Although it was originally thought that the prosodic features only played a subservient role in comprehending an utterance, present research indicates that this was an over-simplified view. In their development of transformational grammar, Chomsky and Halle (1968 p. 25) emphasized the importance of intonation and stress by pointing out that "once the speaker has selected a sentence with a particular syntactic structure and certain lexical items, the choice of stress contour is not a matter subject to further independent decisions." The listener uses the information gained by correctly perceiving the stress contour combined with his knowledge of the language to aid in the identification of the phonemic elements in the utterance.

The prosodic features in natural speech serve many functions, as illustrated, for example, by the word pairs "ob' ject" and "ob ject'" for which the only distinction is one of stress since both contain the same phoneme sequence. In addition, it has been shown that the pitch contour is the major factor in resolving the question-statement dichotomy (Hadding-Koch and Studdert-Kennedy 1964). Daneš (1960), with reference to the syntactic function, states: "in regard to an utterance in isolation, intonation thus integrates the utterance; in regard to connected discourse the intonation delimits the utterances from each other and at the same time segments connected discourse." Recent studies have shown that there is some relationship between the pause pattern of an utterance and: rate of speaking, separation, transitional probabilities between words, level of abstraction, boundaries of utterances, and the characteristics of the speech situation (Boomer and Dittman 1962). Using synthetic speech, Malmberg (1955) found that syllabic division is dependent on the durations of silence with plosives.

Perception of the prosodic feature contours changes the listener's expectation of the kinds of linguistic units to be perceived and

thereby reduces the information load carried by the phonemic elements. For this reason, the prosodic features can enhance comprehension of distorted speech, as well as correct for misarticulations of the speaker. The relationship between the acoustic manifestation of the prosodic features and their relationship to the syntactical and semantic structure of the utterance, can be found in a detailed discussion by Lieberman (1967).

The four parameters which influence the judgment of stress, intensity contour, vocal pitch contour, duration of phonetic and syllabic elements, and the vowel quality, all interact in a complex manner, much of which is not presently understood (Fry 1958). It has been shown that fundamental frequency and envelope amplitudes are the main acoustic correlates of stressed syllables (Lieberman 1960), although other data points to duration as being more important than amplitude (Fry 1955). Examining graphs of intensity as a function of time for natural speech, one finds that they follow a very similar pattern to the pitch contour (Denes 1959) since pitch and sound pressure level are directly related to the subglottal pressure (Ladefoged and McKinney 1963).

Generally, the prosodic features resist distortion and remain basic to the utterance. In analyzing the errors in perception resulting from the removal of the place of articulation feature of the vowels and consonants, but leaving the source and prosodic features intact, Kozhevikov and Chistovich (1965) reported that the word accents associated with the responses were almost always identical to the word accents of the stimulus. As was pointed out in Section 2.5 all the prosodic features, with the exception of vowel quality, which is not very well understood, are spectrally independent and remain unaffected by the transformation.

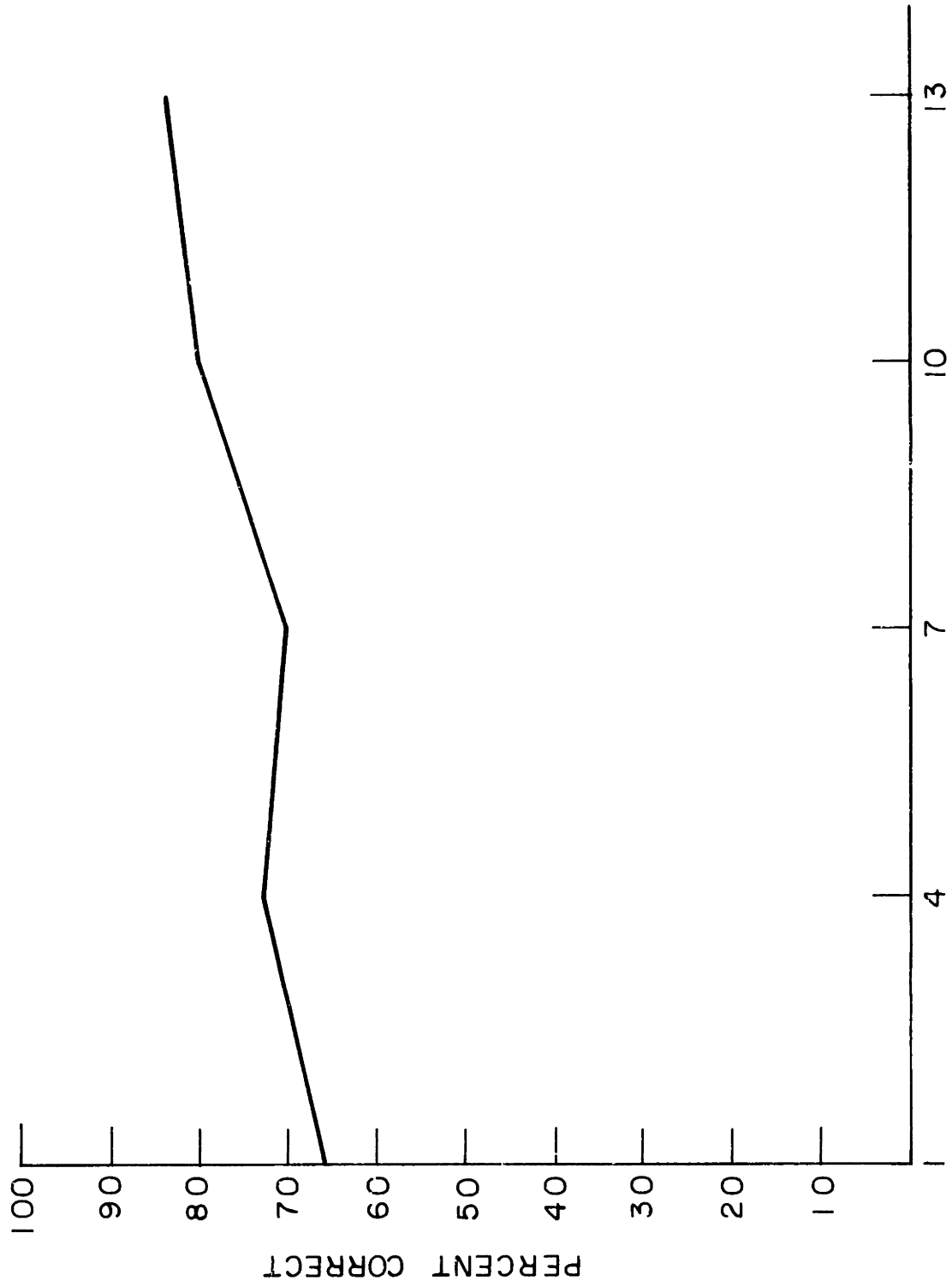
Intonation "is one of the elementary communicative devices of language, forming a special phonological system, serving the organization of utterances differently in different languages" (Danes 1960). The observation that the prosodic feature patterns are a function of

the language, and perhaps to a lesser degree, of individual speakers of the same language, was confirmed in a language discrimination test with spectrally transformed speech. In this test, described in Section 3.31, subjects were requested to judge various passages in different foreign languages as either English or non-English. A simple four sentence passage was read by three native speakers of English and ten native speakers of foreign languages.

The test scores reflect the subjects' perception of the non-phonemic features since he only "hears" overall patterns rather than phonemic sequences. The Kendall coefficient of concordance (Siegel 1956 cht. 9) evaluated for the six times the test was given shows that the relative difficulty, in ranks, for identifying the passages as non-English had a $W = .78$ ($p<.001$). The implication of this high level of correlation between the testing sessions is that the passages possess a characteristic difference in their prosodic features even after spectral transformation. Increased perceptual sensitivity allowed subjects to make more accurate judgments, but the relative perceived "alienness" of the samples remained unchanged. In particular, Arabic, Hindi, Rumanian, Finish and the three English passages were, on the average, scored correctly ($p<.001$); also, French and Japanese were scored significantly above chance ($p<.05$). The scores for Chinese, Russian, German, and Hebrew, however, were not significantly above chance.

Even on the first testing session when subjects had not had any previous exposure to transformed speech they were able to perceive seven of the passages reasonably correctly ($p<.05$). As Graph 4.4a shows the average correct score for all the passages did not improve until session ten. This behavior, in contrast to the other phonetic tests which showed their largest improvement during the first few sessions, corresponds to what will later be described as "break-through". This point corresponds to the time when semantic comprehension suddenly begins to improve. See Section 6.43.

The particular foreign languages that were correctly perceived might be the chance result of the quality of the speakers voice. This



GRAPH 4.4a LANGUAGE DISCRIMINATION ABILITY

was illustrated by the fact that one of the three English passages was significantly more difficult to identify correctly than the others. That speaker had somewhat of a sing-song articulation. The clarity of articulation, rate of speaking, and the intonation patterns peculiar to the personality of the speakers were all uncontrolled variables.

The sensitivity of idiosyncrasies in a speaker's voice was shown during the last session when a special voice identification test was given. The subjects heard the sentence "Joe took father's shoe bench out" spoken by each of the twelve subjects. They were instructed to identify their own and their partner's voices. Of the twelve subjects*, five correctly identified their partner's voice ($p<.001$)**. This clearly shows that subjects perceived an indicative structure in speech which was independent of the linguistic phoneme sequence.

4.5 MODEL OF TRANSFORMED SPEECH

The speech production mechanism has been extensively modeled as a linear system whose characteristics are controlled by the position and movement of the articulators (Stevens 1967). The dynamic behavior of the lips, tongue, mouth, etc., as directed by the higher cortical centers, modulates the carrier source signal with the desired linguistic information. In addition to using the filtering characteristics for generating the acoustic manifestation of the phonemic features, the speaker controls the nature of the source carrier, which contains some phonemic information, such as voicing, frication, or plosion, and the prosodic features.

The resulting speech waveform lends itself to being described by a short-time Fourier transform since this representation extracts

*Only one subject was able to perceive his own voice correctly. This is explained by the fact that one never really listens carefully to one's own voice. Moreover, subjects heard their own voice both transformed and untransformed.

**Assuming that this score is the result of chance the probability of five subjects out of twelve subjects getting the right result is less than .001.

the spectral properties of the oral cavity filtering. The fact that this frequency-amplitude-time pattern is sufficient to generate intelligible speech and that it requires less information than the time representation supports the belief that the spectral properties are a more efficient speech characterization. Bandwidth compression systems and vocoders use this fact (Schroeder 1966).

Since it was shown that listeners perceived both the source and prosodic features of transformed speech correctly, one can think of the transformation as only affecting the filtering characteristics of the oral cavity. In essence, a human speaker could produce spectrally rotated speech which would be equivalent to that created by the system, at least from a perceptual point of view, if he could move his articulators so as to produce the right filtering properties. Although spectrally rotated speech is probably not, in fact, humanly reproducible, one can postulate a speaker whose mouth was suitably distorted such that he could generate the appropriate geometry necessary to simulate the effects of the system.

Modeled in this way, spectral transformation is a unique speech defect which is, perhaps, analogous, in terms of perceptual difficulty, to the speech produced by someone with a cleft palate or to that of the congenitally deaf. It is well known that those who live with people with speech defects often learn to understand them quite well even if others not familiar with their speech patterns cannot. A mother understanding a child's baby talk is a similar situation. Only if the defect produced a distortion which destroyed most of the linguistic information would one predict that the spectrally transformed speech would not be learnable. The various experiments described in the previous sections have shown that the major loss in information occurs with the place of articulation feature and the identification of the lax vowels.

Linguistically, the confusions resulting from spectral transformation have very little effect on comprehension. Miller and Nicely (1955) made the interesting observation that listeners could understand speech in which one phoneme was used for all the members of the same

source feature class. That is, whenever a /p/, /t/, or /k/ occurred, a /t/ was spoken and similarly for the other four source-feature groups of sounds. This speech, called "elliptic", was understandable although it sounded as though the speaker had a marked dialect or speech defect. Denes (1963) pointed out that the basis for this phenomenon is that the place of articulation feature carries very little information. Only 25% of the consonant minimal pairs* contain just a place of articulation feature difference. Of the 12 most frequently occurring minimal pairs, only one has a place of articulation feature difference.

Furthermore, when such minimal pairs occur they can often be resolved on the basis of the probability of occurrence. Since some phonemes are much more prevalent, pure guessing of the place of articulation feature results in the listener being right over 55% of the time. The phonemes /t/, /d/, /n/, /s/, /z/, and /l/ all have the alveolar place feature and occur more often than all other consonants (Tobais 1959). The probability of guessing increases still further when one includes the second and higher order statistical properties of English. Shannon (1951), in his theoretical treatment of the entropy in printed English, which is presumable analogous to phonemes in spoken English, proved that the information conveyed by one extra letter in a long sequence is less than one bit on the average; whereas, the information conveyed by a single letter in isolation is almost five bits. The factor of five is a measure of the redundancy and means that, in theory, one need only transmit one out of every five letters, using the right encoding, in order to recreate the original message.

These observations in conjunction with the experimental data in the previous sections lead one directly to the assumption that spectrally rotated speech is understandable. The interesting question yet to be answered is why comprehension of transformed speech generally begins after about six sessions and not immediately, even though the

*Minimal pair is two words which differ in only one feature of one of the consonants. For example, /bɛd/ and /dɛd/ are minimal pairs in the place of articulation feature.

phonemic perception showes the greatest improvement during the first few sessions.

4.6 DIFFERENTIAL ABILITY OF SUBJECTS

The innate ability of all subjects to master transformed speech, ignoring factors of personality, effectiveness of learning strategies, and motivation, is not expected to be the same. In the foreign language case, this was explained by Carroll in the following way: "There are many circumstances (e.g. intensive courses for government personnel) where it can be taken for granted that motivation is uniformly high and where the instruction is of very good quality; even in these situations, success varies widely. Further, these individual differences are very hard to modify; Docus, Mount and Jones found high correlations between early and final examinations in the eight intensive courses of the Army Language School and inferred from this that 'important factors do exist for the prediction of language proficiency.'" (Titone 1964, pp. 29-30).

The learning curves included in the previous sections show average performance data which obscures some very pertinent information contained in the individual scores. The scores for some of the tests ranged from almost 0% to 100%, while the range of scores on the other tests was very narrow.Awide range should not be interpreted as meaning that some subjects did not pay attention, since no subject did poorly on every kind of test; rather, the range should be interpreted as being a reflection of different kinds of cognitive abilities.

The differential ability of the subjects can be used to investigate "modes of perception" for different speech tasks. It may be hypothesized that if two types of tests measure responses from perceptually related processing then the relative performance of subjects on the two tests should correlate. Or, if they do not correlate, then the responses may be said to be based on different cognitive mechanisms or different manifestations of the same mechanism. In other words, it might be found that those subjects who do well on task "A", such as learning German, also do well on task "B", such as learning French, in

which case the psychological mechanisms used to learn these two tasks are related. Alternatively, if task "A" was memorizing a long poem and task "B" was learning calculus, then one might find that the relative performance on the two tasks does not correlate. In terms of this example, one could say that the learning mechanisms brought into play with French and German were more "similar" than those for memorizing poems and learning calculus. There are many weaknesses in such a pragmatic definition of cognitive similarity, but it suffices to demonstrate that at the feature level there are many modes used in the perception of speech and that these modes may be relatively independent.

On each test every subject is assigned a comparative rank score which replaces his absolute score. In this way, the effects of average learning are removed and variations in the tests no longer manifest themselves. The rank measure is completely relative and therefore requires the use of non-parametric statistics (Siegel 1956). Before any correlation studies can be made it is first necessary to determine the stability and consistency of the subjects' rank for one task over the course of the entire experiment. Computation with the Kendal Coefficient of Concordance, W, shows that the rank ordering for a particular test was highly correlated with similar tests given during other sessions. That is, a subject who, for example, was able to identify vowels better than other subjects on session three was very likely to do so on sessions six, nine, etc. As can be seen in Table 4.6a, W is typically greater than .6 with a level of significant less than .001. Thus, whatever a test was measuring, it measured it consistently; and whoever did well once usually did well again. Because learning took place, a high W also says that a subject who did well initially was likely to learn as rapidly as, or more rapidly than, others.

The subject's average rank scores for a particular type of test is considered to be the best measure of the subject's ability since the average scores form the basis for the Kendal correlation test. In other words, the overall average measure of a subject's ability to perceive, for example, vowels is the average of his rank scores on each of the vowels tests (of the same type). In the remaining correlation studies,

the average ranks for each type of test are considered raw data and then re-ranked. Using this new set of ranks for each of the test types listed in Table 4.6a, the Spearman correlation between the tasks can be determined.

It is found that the four consonant task listed in Group I are all highly correlated with each other, but are not correlated with any of the other tasks. Similarly, none of the vowel tasks in Group III correlate with any other task except those within the group. The same is true for Group II, consonants. The fact that the ABX consonant discrimination ability is not related to consonant identification but is related to vowel discrimination is of great consequence. This suggests that a sensitive listener who can detect minute differences in sounds may not, necessarily, be better able to identify phonemes. Equally important is the observation that the Spearman coefficient between correct identification of a phoneme and the correct perception of its place of articulation is .95 and .92 for vowels and consonants respectively; but the correlation between place of articulation for vowels and consonants is not statistically significant. Both are spectral features yet they do not seem to be perceptually similar. That is, the subject who can perceive the place of identification correctly for the consonants may not necessarily be able to do so for the vowels, and vice versa.

In order to assess the relationship betwen the four groups, the Kendal W correlation test is used to verify that the rank ordering of each of the tasks listed in each group is very highly correlated. The results, shown in Table 4.6b, indicate that within a group the rank ordering is very similar. As before, a new set of ranks for each of the groups is created by re-ranking the average ranks of the tests included within each group. The resulting set of four ranks for vowels, consonants, discrimination and language, are again correlated with each other. Not only was there no significant correlation between them, but they approached 0. in almost every case!

	TASK	W	<p*	
1.	Consonant identification	.669	.001	
2.	Voicing-nonvoicing feature	.439	.01	Group I
3.	Mode of articulation, consonants	.394	.01	
4.	Place of articulation, consonants	.684	.001	
5.	ABX consonant discrimination	.512	.001	
6.	ABX unvoiced plosive discrimination	.606	.01	Group II
7.	ABX vowel discrimination	.672	.05	
8.	Vowel identification	.690	.001	
9.	Tense-lax feature, vowels	.679	.001	Group III
10.	Place of articulation, vowels	.594	.001	
11.	Vowel pair recognition	.692	.001	
12.	Language identification	.330	.05	Group IV
**	Unvoiced plosive identification	.231	--	

TABLE 4.6a

KENDAL W CORRELATION COEFFICIENT FOR TESTS

*the level of significance is a function of both W and the number of tests included.

**Since the raw scores for these tests were only slightly above chance there were no correlations between them. For this reason, average ranks have no meaning and the test results are not included in the correlation studies.

	GROUP	W	<p-
I	Consonants	.664	.01
II	ABX discrimination	.841	.01
III	Vowels	.809	.001

Table 4.6b Kendal W coefficients for groups listed in Table 4.6a.

Although the experimental and statistical controls are not adequate to conclusively demonstrate that identifying consonants, identifying vowels, discriminating phonemes, and discriminating prosodic features are independent tasks, there is a strong suggestion that this, in fact, is the case. Moreover, this observation is an extension of the results from the experiment with left- and right-ear dominance for vowels and consonants respectively (Shankweiler and Studdert-Kennedy 1967). They hypothesized that consonants should be perceived better by the right ear since the left hemisphere of the cortex contains the language center. The results from the transformed speech data, however, suggest that the actual cognitive mechanisms may be different for vowels, consonants, etc. The implication is that different listeners may use different kinds of cues for perception of speech depending on which abilities function best. A more careful experiment would be needed to answer this question however.

CHAPTER V

COMPREHENSION OF TRANSFORMED SPEECH

The comprehension of words and sentences, unlike the perception of phonemes, involves the extraction of linguistic meaning from the acoustic wave. Even though the manifestation of words is often thought of in terms of the phonemic segments which make them up, there is no conclusive evidence to indicate that the "whole is the sum of the parts". There is, in fact, evidence to the contrary. Liberman (1967) makes the very succinct point that the duration of phonemes is too short to allow for the auditory system to perceive them as units. Perception must, therefore, be based on larger elements.

The currently accepted theories of speech perception all view cognition as an active process in which a verbal hypothesis is constructed from the listener's knowledge of the language, the environmental bias, and some aspects of the acoustic signal, rather than viewing perception as a direct result of the stimulus (Chomsky and Halle 1968, p. 294). Because perception is not merely the result of the stimulus, the listener can correctly perceive the word or idea without actually having "heard" it. Some of the factors which may contribute to comprehension, in one way or another, are the semantic context, the syntactic structure, emotionally biased stress patterns, the probability of a word's occurrence, the environmental context in which the utterance was spoken, and the predisposition of the listener.

5.1 EFFECTS OF CONTEXT AND REDUNDANCY

There are two factors which have been shown to govern the intelligibility of a stimulus word in the presence of noise: the probability of its occurrence, and the number of syllables. When a word is heard in a normal sentence, its probability is a function of both its frequency of occurrence in normal English and the probability that such a word is likely to fit into the surrounding semantic and syntactic context. Miller, Heisse and Licten (1951) showed that with a given signal-

to-noise ratio a word, perceived correctly 85% of the time in a sentence, would be heard correctly 55% of the time when presented in isolation. The use of the word in the sentence changes its probability since most words could not be used and still have the sentence be meaningful.

Equivalently, the probability of a correct response for words in isolation may be increased by decreasing the number of acceptable responses in the set. In the same experiment, Miller, Heise, and Lichten found that with a low signal-to-noise ratio digits were almost always heard correctly even though words in sentences were completely unrecognizable. The digits form a very small set of possible choices and therefore have a comparably high probability of occurrence. Confirmation of this phenomenon was found by Pollack, Rubenstein and Decker (1959) in an experiment which showed that the effect of frequency of occurrence could be removed by having subjects learn all the elements in the set of stimulus words. Knowing the elements in a small closed set appears to equalize the relative probabilities.

In contrast to what the frequency of occurrence phenomenon would predict, long, uncommon words are often more intelligible than shorter, more common words since, as Savin (1963) points out, a multiple-syllable word carries its own context. The noise threshold for words of three or more syllables, according to an unpublished experiment of Miller, is about the same as for one-syllable words in sentences.

Since a rare, polysyllabic word would never be confused with a one syllable-word, probability can not be the only factor in determining confusions. That is, confusions are between words which are acoustically similar, having for example the same number of syllables or the same consonant-vowel pattern. Whether an indistinctly heard word is identified correctly depends upon the realtive frequency of that choice within its own confusion class. For this reason, a multiple syllable word, although infrequently used, is usually not sufficiently similar to any other word which it could be confused with. The Miller

and Nicely (1955) experiments suggest that confusions, with [illegible] masking, would be between those words which differ only in the p[illegible] articulation feature of the consonants.

Speech communication involving utterances larger than one word symbols, such as sentences, is a highly complex behavior little of which is currently understood. One theoretical formulation of psycholinguistics identifies two levels of linguistic representation and a generative grammar to translate between them. It is felt that the idea or semantic reality manifests itself in the "deep structure" and is the received or transmitted content of the utterance. For the speaker, the generative grammar operates on the message content and transforms it, according to the linguistic rules of that language, into a "surface structure" representation. The surface structure, being implicitly closer to the actual utterance, is then transformed into the phoneme sequence. For the listener, the process is reversed; the grammar operates on the phonetically decoded surface structure transforming it to a semantic deep structure. A more detailed exposition of this theory can be found in Chomsky (1964), Chomsky (1966), and Garrett and Fodor (1968).

This formulation allows for two kinds of redundancy to be isolated. The generative grammar specifies that only certain combinations of content and formative words are to be considered legal utterances. That is, for example, "You him book gave" would not be allowed, but "You gave him a book" would be. This kind of structural redundancy limits the choice of word sequences which one would expect to find, which, in terms of information content, increases the probability of a particular word appearing and decreases the probability of others once part of a sentence is perceived. Equally, at the deep structure, only some utterances, by virtue of their meaningfulness, would be accepted. Although "colorless green ideas sleep furiously" satisfies the syntactic encoding rules, it does not yield any acceptable semantic content. This kind of redundancy is also important in conversations composed of many sentences with a common theme. Since succeeding sentences are related to each other by their content, there is a predictive component

which allows a listener to guess much of each sentence.

This formulation is only schematic and should not be taken too literally since there is strong interaction between the semantic, syntactic and phonemic levels of perception. Numerous experiments illustrate this point as does the more commonplace observation that a listener can often paraphrase the meaning of an utterance without being able to perceive or recall the exact word and phoneme sequence. Only a few aspects of the deep and surface structures have as yet been explored, but these experiments tend to be encouraging although quite inadequate.

Boomer (1965) found, in analyzing normal utterances, that hesitation pauses occurred predominantly at the boundaries between "phonemic clauses". Similarly, Garrett (1968) stated that the perception of artificially induced pauses was much smaller at the phrase boundaries than elsewhere in the sentence, thus showing that the listener expects a pause as a natural part of the utterance. In a related set of experiments, Garrett, Bever, Fodor (1966) and Fodor and Bever (1965) found that clicks presented dichotically to subjects listening to sentences were reported at the constituent breaks. This has given support to the hypothesis that perception occurs in large sections and that only between sections is the perceptual system free to perceive the extraneous noise. Further proof of linguistically grouped sections was shown by Huggins (1964) in an experiment involving the switching of the speech signal between the two ears. He found that the critical switching rate was a function of the speed of presentation and not an inherent scanning time in the perceptual system.

The deep and surface structures of an utterance can significantly affect perception. Johnson (1965) demonstrated in a paired association test that the probability of a transitional error, i.e. incorrect perception of a word following a correctly perceived word, was less when the two words were within a phrase unit and greater when they were across it. It has also been shown that learning and recall are affected by the surface structure (Kent 1963) and by deep structure (Miller and

McKean 1964; Mehler 1963). Apparently the semantic and syntactic redundancy shifts the relative probabilities and thereby affects perception. In the presence of noise, syntactically correct nonsense sentences are more difficult to perceive correctly since they lack a predicting semantic content. Pseudo-sentences which lack both semantic and syntactic structure are still more difficult to perceive (Miller and Isard 1963). Such sentences cannot be produced with the syntactically related prosodic features and one would therefore expect them to be much less intelligible in the presence of noise. Similarly, the same situation occurs for the recall of sentences (Marks and Miller 1964).

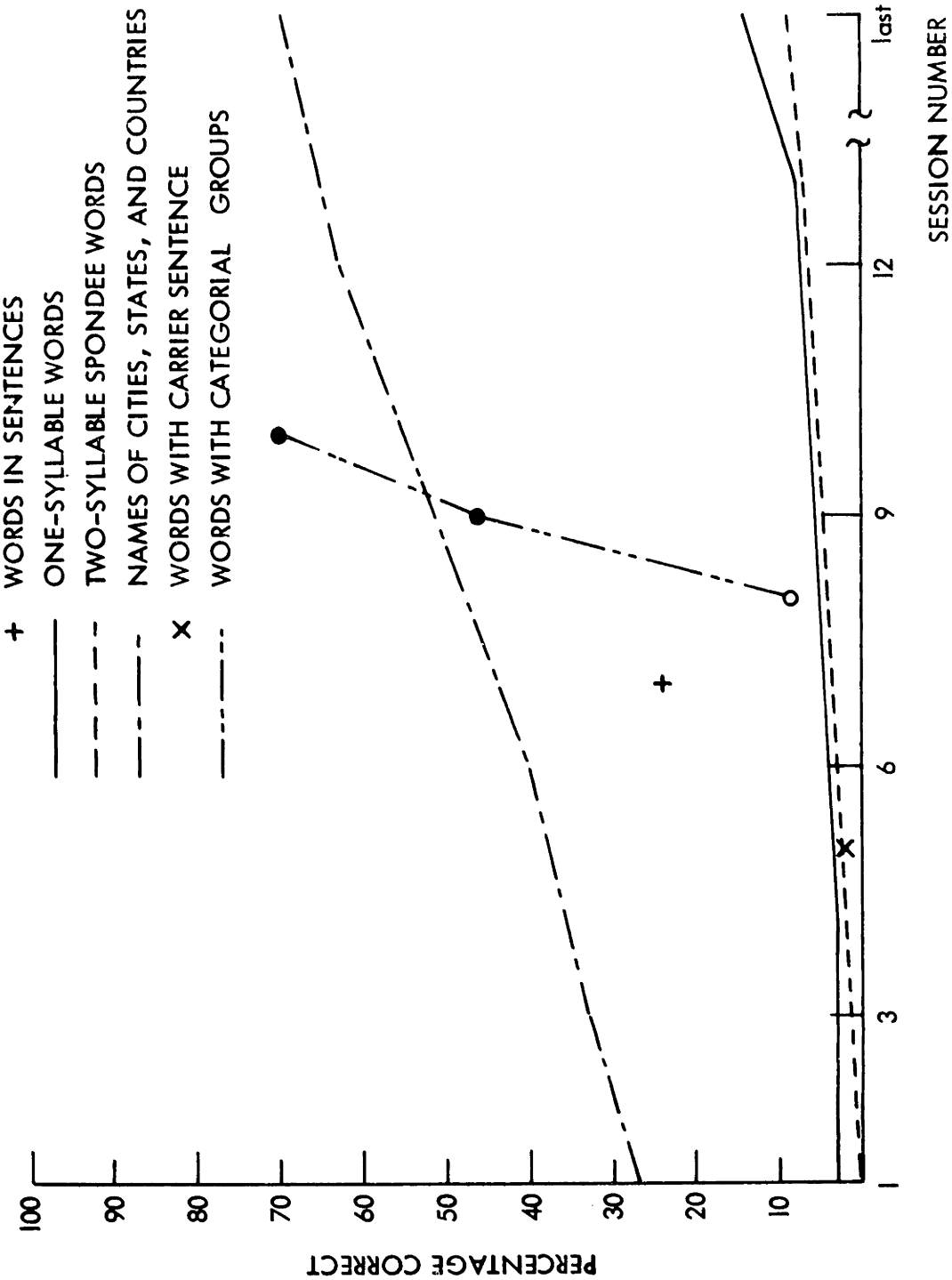
These various experiments do not actually give much insight into the psycholinguistics of perception, but they do give evidence for the existence of a non-acoustic cognitive processing which affects the basis of verbal perception.

5.2 COMPREHENSION OF WORDS

5.21 OPEN-SET LISTS

One measure of transformed speech comprehension is the ability to correctly perceive isolated words with no contextual cues. Such words are only slightly easier to identify than random nonsense syllables or isolated phonemes. Following some of the conversational practice sessions, subjects were given tests composed of either one- or two-syllable word lists taken from the Harvard PB (phonetically balanced) lists (Egan 1948). See Section 3.36 for details about the test content and methods. The two-syllable lists were composed of "spondees", such as blackboard and horseshoe, since these homogeneous words have the property that they all reach the threshold of hearing within a very narrow range. Although these lists have sometimes been criticized for not being adequately representative of English, these tests only served to give a general measure of comprehension.

As can be seen from the results in Fig. 5.2a, subjects found it very difficult to identify these words correctly and the average



GRAPH 5.2a WORD COMPREHENSION.

score, even by the end of the testing series, for one- and two-syllable words was only 14% and 9% respectively. The two-syllable words were as difficult, or more difficult, to perceive correctly than the one-syllable words. Although this unexpected result would tend to minimize the importance of internal redundance of long words, the two-syllable spondees were rather uncommon compared to other two-syllable words with similar phonemic structure. Had they been typical, the score for two-syllable words would most likely have been higher.

The curves in Graph 5.2a for one-and two-syllable words obscure the fact that almost all the correct responses resulted from one pair of subjects. This pair, who incidentally also had the most fluent conversations, scored an average of 17% and 45% for one- and two-syllable words respectively on the last session. In contrast to the other subjects, they learned to recognize isolated words and they found two-syllable words much easier to identify. The reason why one pair of subjects should show a very pronounced learning of isolated words and others should not is not apparent. Their scores on the phonemic tests were generally good but not exceptional. Only a pair of twins, who were used in the pilot study, showed the same ability to perceive words in isolation.

The difficulty in perceiving isolated words is revealed by examining the error patterns in the responses. In general, the incorrect responses resembled the stimulus word in several ways; the consonant vowel pattern was preserved, the source features of the consonants were correct, and the accent was on the right syllable for the two-syllable words. Only the place of articulation feature of the vowels and consonants was readily confused. For example, "rub" was heard as "red", "plush" as "class", "pants" as "cards", and "baseball" as "dusty". By accident, a two-syllable word was included on one of the one-syllable word tests. Subjects responded to this word either by making a note that the word had two syllables or by responding with a two-syllable word. It thus appears that subjects wrote down the first word that came to mind which satisfied those featues which they perceived. If the stimulus word was the first to emerge into the subjects'

consciousness, then the response would be correct. For the stimulus word "armchair", a majority of subjects responded "answer" because it is phonetically similar and the subjects were possibly "thinking", in a sense, about the notion of "answering the test".

Since it was found that the perception of vowels could be influenced by a carrier sentence (Section 4.24), in a test given on session eight, each stimulus word was preceded by a sentence acting as an acoustic context. Even with this context, no improvement in recognition was observed. Moreover, not only was no consistent patterns of vowel confusions observed on this test, but no pattern emerged on any of the word tests.

The consonants appear to be the primary factor in determining the response word, with the vowels being subsidiary. Huey (1968, p. 214) reports that in primitive written languages the vowels were an implied part of the consonants acting as a necessary element for the pronounciation of the consonants. Even in modern Hebrew the vowels are only sometimes included in written representations. The differences in the dialects of English are primarily shifts and modifications of vowel phonemes. The instability of vowel perception, as discussed in Section 4.24, implies that once the listener had "perceived" a response word which satisfied the consonant and syllabic pattern he could "force" the vowel sound to the appropriate phoneme.

The fluid nature of the vowels is not analogous to the confusions of consonants within the same source feature class, since the vowels, unlike the consonants, did show consistent behavior in the phoneme tests and were readily discriminable in the ABX tests. Thus, the lack of a pattern of confusions for vowels in the word tests should not be attributed to spectral transformation. Rather, it must reflect some manifestation of the cognitive process. Paradoxically, by the end of the experiment most subjects could converse with each other yet they could not identify words.

5.22 CONTEXT AND SET-LIMITED LISTS

Several other word tests, discussed in Section 3.36, were given throughout the experiment to measure comprehension of set-limited words. In one series of tests, repeated throughout the experiment, the subjects were told that the stimulus words would each be the name of a city, state or country. Using such a set has the advantage that the subjects were throughly familiar with the acceptable stimuli and yet the set contains at least several hundred elements. Quite unexpectedly, subjects scored very high, starting at 27% on the first session and improving to 70% on the last session. The learning curve, shown in Graph 5.2a, is monotonic and shows no indication of being asymptotic to a value less than 100%. The additional category cue appears to be a very significant factor in the high level of performance, especially when compared to the open-set stimuli word tests.

The error patterns for this names test are somewhat different from that of the one- and two-syllable word test. The number of syllables and accent were a more important criterion than the actual phonemic composition. For example, "Switzerland" was heard as "Washington", "Washington" as "Newfoundland", "Ireland" as "Oregon", and "Czechoslovakia" as "Cincinnati". Most of the incorrect response words contained the same number of syllables and a very similar rhythm. Another example was "Boston" which was heard as "Camden", "Athens", "Hampton", "Easton", "London", and "Houston". In this example, the last syllable is similar phonemically to the stimulus word, but the first syllable appears almost random.

Another testing series, designed especially to verify the effect of limited-set words, was given on sessions eight, nine and ten. Four groups of ten words composed of colors, parts of the body, furniture, and vegetables were selected. On session eight, a representative sample of 20 words, five from each group, were presented to the subjects with no instruction other than "write down the word spoken". The four groups were also presented on the next two sessions, but this time they were told "the next ten words will all be ....(colors)...". When

the subjects were not told the category name for the stimulus words the average score was 8% which was approximately the same score for the one- and two-syllable words. However, when the names of the categories were given the average score was 58%. This score is approximately the same as that of the names of cities, states and countries test. Limiting the range of possible acceptable responses restricts the confusion class and generally results in there being only one word with the correct phonemic structure. The erronious responses, therefore, had almost no phonemic resemblence to the stimulus word. "Cabbage", for example, was heard as "turnips", "potato ", and "spinach".

The perception of a stimulus word in a known sentence also shows the same phenomenon. The surrounding semantic context of the sentence acts to restrict the number of acceptable choices for the unknown word in the same way that category names focus the listener's attention on a particular class of words. The average score on this test, given on session seven, was only 25%. Although this score is significantly above that of the open-set word tests, it is not as high as the score on the names of cities, states and countries test. Again, it was found that subjects tended to rely more on what would make sense in the given context than on what the word sounded like. In the sentence "Read the verse aloud for ....", the correct answer was "pleasure", but the responses included "error", "mother", "color", "practice" and "tune". The responses were always semantically and syntactically correct even though the subject had to "stretch" his perception of the phoneme sequence to fit the word he thought ought to be there. The relationship between stimulus and response in a complex test is quite ambiguous.

The subjects reported, upon being questioned, that they thought they heard the word which they wrote down even though it might have been wrong. This introspection emphasizes the point that the listener is not consciously guessing a response which he thinks is close, but is "hearing" the word which first appears in his consciousness. This

is, perhaps, illustrated by the following examples, "France" was often heard as "Paris", "England" as "London", and "Paris" as "France". Several different names were heard as "Hanoi" which at the time of the test was a current place in the news media. The phonemic and syllabic differencesare sufficiently great so as to preclude the possibility that the responses were only based on phonemic confusions.

Skinner (1936) has demonstrated that indistinct speech-like stimuli can act as an auditory Rorschach test in which the listener free-associates to a response. In his experiment with a sequence of vowel like sounds near the threshold of hearing he found that subjects "heard" sentences and were often convinced that their response was correct. One subject reported:"Funny, they sound like nothing at all until suddenly they sound like something. I listen. It starts saying something. The more I listen, the more its says that one thing." This illustrates the ability of the cognitive system to impress a meaning (percept) on a stimulus to organize its randomness. Skinner found that the content of the responses was sometimes related to objects in the immediate environment. A clock striking the hour led to a response "half-past", an automobile horn outside the room led to the response "automobile", and many subjects frequently heard "How do you do". The same situation was found with transformed speech. On the two-syllable tests the responses "bastard", "practice", "annoy" and "testing" were heard.

5.3 SENTENCE COMPREHENSION

5.31 CONTENT INDEPENDENT SENTENCES

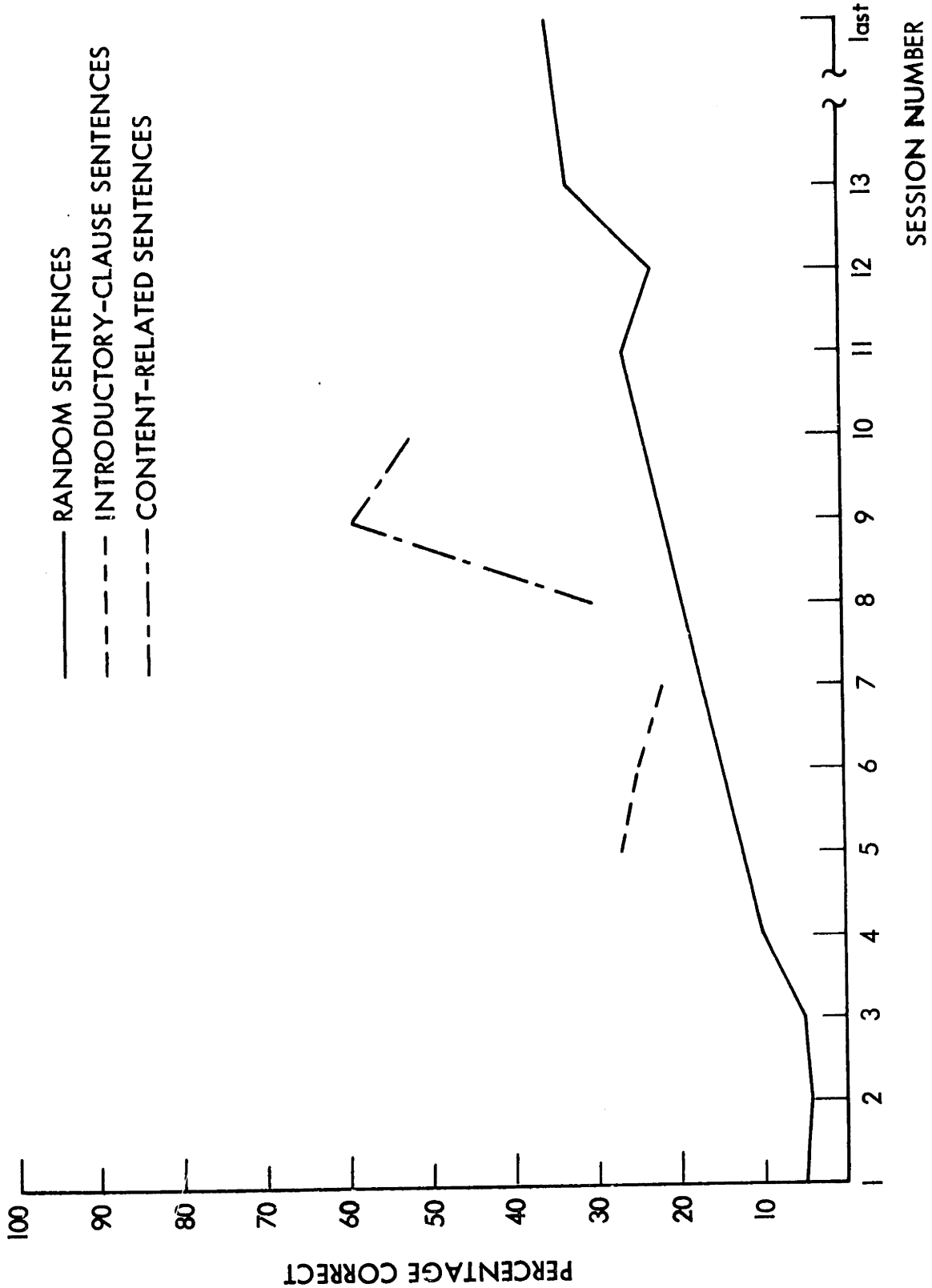
The comprehension of whole sentences, in contrast to the perception of isolated words, allows, or perhaps requires, cognitive processing of the syntactic structure and the incorporation of the semantic relationships. As a measure of comprehending transformed speech, perceiving sentences has the advantage of being a task which is closely related to normal conversation, and has the disadvantage of having a complexity which is beyond our present knowledge of the psycholinguistic influences on verbal cognition.

Following the conversational practice with transformed speech, one of three kinds of sentences tests, described in Section 3.37, was given to the subjects. The first group contained sets of ten unrelated sentences with an average length of eight words. The sentences in the second group all contained an introductory clause or phrase, given to the subjects in the instructions, and an unknown main clause with an average length of about six words. In the third group, all the sentences were related by a common theme and formed a simple story.

Although the inherent complexity of a sentence is an appropriate attribute for a semantic stimulus designed to evaluate comprehension, this same complexity makes it difficult or impossible to control for relative difficulty. Based on present knowledge, two sentences which are equally difficult to comprehend under some adverse condition, such as additive noise, might very likely not be equivalent under the condition of spectral transformation. The only true comparative measure, then, is the differences in the score for the same sentences used at different times during the experiment. The sentence tests, which for many reasons were not well designed, were repeated during the second half of the testing series. The tests given on sessions one through ten were repeated on sessions eleven through twenty. As a result of an oversight in the experimental design, the sentence tests from the three groups were not interleaved, rather they were given in sequence. Group I was given on sessions one through four, eleven through fourteen and also on the last session; Group II was given on sessions five through seven and fourteen through sixteen; Group III was given on sessions eight through ten and seventeen through nineteen. The curves in Graph 5.3a, therefore, do not give a continuous measure of performance.

It can be seen in Graph 5.3a that the average comprehension on the random sentence tests improved from 5%* on the first session to 35% on the last session. The inherent learnability of transformed

*In this set of tests, the scores are based on the percentage of the syllables scored correctly. Thus, the sentence "The boy had stopped at the car." contains seven syllables, and the response "The boy was stopping at the car." is scored as 75%.



GRAPH 5.3a SENTENCE COMPREHENSION.

speech is further demonstrated by the fact that one pair of subjects, who also scored very well on the word tests, were able to comprehend every sentence correctly on the last session. The learning curves, suggest that, with enough conversational practice, subjects could learn to understand sentences without having any contextual cues.

A very profound phenomenon was observed in the subjects' responses, mostly during the initial sessions; subjects were unable to comprehend the content, meaning, or the individual words in any sentence, but they were able to perceive, often completely, the syntactic structure of the utterance. The 5% score for session one obscures the fact that the only words identified correctly were function words, which, by themselves, contained no meaningful semantic information. The responses for one typical sentence, listed below, illustrate this point clearly.

a)	The	is	
b)	Los Angeles	has	and trees.
c)	The	has	and praise.
d)	The	has	and pies.
e)	The	have	
f)	The baker	has little	and pies.
g)			and

Responses to test sentence "The farmer has many chickens and cows.

Other examples of perceiving phrase structure were also found, such as "into the sky", which was heard as "under the stairs", "under the square", "under the scale". All the initial "the" articles were perceived correctly as were many of the medial "and", "is", etc. The most striking example of the phenomenon was the sentence "Hoist the load to your left shoulder." which was heard as "Turn the page to the next lesson." In this example, the corresponding words in the response have the same part of speech, or function, as in the original stimulus sentence; yet the meaning of the response sentence has absolutely nothing

to do with the stimulus. Even during the first session, when subjects had had no conversational practice, their responses, even though sparce, were almost always correct in syntactic and phrase structure.

The implications of this unexpected result are two-fold. Firstly, by being able to identify a single word in the continuous unsegmented speech signal, the subjects were using some aspect of acoustically manifest prosodic features to segment connected speech. There is no possibility that segmentation was the result of hearing a sequence of words since subjects were not able to identify any of the content words. Based on the confusions of consonants, one might have expected that "cows and" would sometimes be perceived as "go sand"; however, this kind of segmentation never occurs, even though there is no pause between "cows" and "and".

Secondly, perception of the syntactic structure suggests, if not proves, that there is an acoustic manifestation of prosodic features related to the surface structure. Although it is not clear where and how the correct features are being manifest, it does demonstrate that structural coding is an independent domain apart from the content and phonemic structure of the utterance. Furthermore, the coding of the surface structure seems to be the basis for a subject's ability to segment words. The ease with which the subjects ignore phonemic matching as a basis for judgment but demand stress matching would imply that the unified stress pattern can have a higher perceptual priority than the phonemic sequence. It might be that the prosodic structure of an utterance is the medium upon which the phonemic and word units are "attached", in the same way that a canvas is used to hold the paint for a picture.

The error patterns on the sentence tests began to change as the subjects gained more experience with transformed speech. The responses on sessions two and three were fewer and more cautious than on session one. Although this might be the result of those sentences being more difficult, it is felt that the subjects, after practicing for a half hour, were more unsure of their perception since they found

themselves unable to communicate with their partners. Whereas on testing session one, subjects received no feedback about their accuracy and could have assumed that they were performing reasonably well.

By session four, the frequency of response words was again up to the level of the first session. The increased score for this session reflects the increase in correctly perceived content words, in addition to the function words. Almost without exception, comprehension of a content word occurred with the phrase in which it was embedded. For example, "into the city" and "to understand it" were scored correctly, but there were no examples of content words without phase units.

On sessions five, six, and seven, sentences with a common initial clause or phrase were given in an effort to determine if an acoustic context would aid comprehension. This possibility was suggested by an experiment of Pollack and Pickett (1964) which showed that the presentation of a long speech utterance previous to the stimulus word could enhance its perception even though the subjects were told in advance the content of the context. The results using this kind of test with transformed speech, shown in Graph 5.3a, indicate a significant improvement. However, the results are somewhat misleading since the existence of an initial subordinate clause defines, to a greater or lesser extent, the range of structure and content which can follow. For this reason, it is difficult to decide if the improvement is the result of acoustic context or the reduction in semantic and syntactic ambiguity. This test, being poorly designed, does not really resolve any of the issues raised by the sentences in Group I.

Paradoxically, the higher score on the sentences of sessions eleven, twelve and thirteen results from improved recognition of content words, but, in addition, the syntactic structure of the responses was no longer as similar to the stimulus structure as it had been on session one. It appears that subjects were now willing to depend more on the content words, which they thought they could perceive, and less on the syntactically related stress pattern. The subject who was able to perceive the exact structure in the example "Turn the page to the

next lesson" had a rank score of 11 out of 12. That is, he perceived the function words clearly, but was never able to advance to the content words. Thus, it might be hypothesized that if a subject perceived a group of words which related to each other in a meaningful way, then he would generate a structure into which they could fit and reduce his dependence on the perceived structure. The role of prosodic features is not at all clear.

5.32 CONTENT RELATED SENTENCES

The suggestion that semantic content words could over-rule syntax in the perception of spectrally transformed sentences was tested in sessions eight, nine, and ten. Tests, composed of ten sentences forming a conversation on one theme, were presented to the subjects with instructions informing them of the theme. The actual test sentences are listed in Appendix A for reference. The algorithm for scoring the responses was almost identical to that used in the random sentences; that is, a two-syllable word was worth twice as much as a one-syllable word.

Knowing the general content of the sentences enhances perception very significantly. As can be seen from Graph 5.3a, the scores for this groups of tests were more than twice the score for the random sentences. Once the listener has a reasonably accurate notion about the kinds of sentences which ought to appear, perception of those sentences is enhanced. The contextual semantic information acts to limit the set of possible words in very much the same way that perception of words is improved when the subjects were told the category of the words. In some cases the subjects were so convinced that a particularly appropriate word should be there that they were willing to "bend" the syntactic and phonemic structure to make it fit. For example, the instructions of session ten told the subjects that the sentences were taken from a conversation between a young couple and a travel agent; the following were the responses to the second sentence,

"I have a special two-week trip to Europe."

I had expected to take a trip to Europe.
I have especially for you.....
.....and here we'll take.....
I have a special preview trip to Europe.
I have a special .....packed.
I have a special travel plan for you.
I have especially for you a trip to Hawaii.
I have a special two-week trip to Europe.
I have a travel plan for you.
I have a special trip for you.

Another particularly good example is found in the test of session eight which contained the instruction telling the subjects that the sentences were a conversation between two students on a Friday afternoon. The following were the responses to the seventh sentence:

"Do you like taking so many courses?"

Will you have problems in your classes?
Do you have trouble with so many classes?
Do you like taking .....classes?
.....haven't taken classes?
Do you have problems.....classes?
Do you have problems.....?
Do you have any trouble in classes?
Do we.....in our classes?

Each of the responses, or partial responses, is quite reasonable when considered in terms of the context created by the previous sentence, "This one will be difficult since I have not studied." It might be suggested that there is more than a casual relationship between the word "difficult" in the previous sentence and the words "problems" and "trouble" in the response sentences. Upon hearing "difficult" subjects had become prepared for a word or idea which would relate, or be associated, to it. This contextual bias induced the perception of a sentence

which was not directly related to the actual stimulus sentence. That is, the semantic unity of a conversation makes the sentence "Do you like taking so many courses?" seem unlikely since it does not follow from the previous sentence.

Another significant aspect of these groups of sentences is that they, unlike the random sentence groups, have much higher scores for the first five sentences than for the last five ($p<.025$). This might be the result of the inter-relationship between the sentences, in that not understanding one sentence makes perception of the next one more difficult. In several of the tests the topic being discussed gradually changes from one aspect of a theme to another. "Loosing track" would impair perception. In contrast, the comprehension of the random sentences increased towards the end of each test, as a result, perhaps, of adapting to the author's voice and speaking style.

Other experimenters, however, have found similar kinds of phenomena occuring.

> "D. B. Fry has given a remarkable demonstration of the way in which a priori knowledge bears upon recognition. He has made a gramaphone recording of two men holding a conversation, but with their speech so artificially distorted that not a word can be recognized. After one playing of the record, the listener is informed that the speakers are discussing the subject of buying a new suit; they refer to their tailors, the price of clothes, styles, et cetera. The record is then played a second time, and most listeners are able to follow the conversation. The words jump out at one." (Cherry 1957, p. 276)

Similarly, Tonndorf (DB 1968) tells about one pilot who was "deaf as a doorpost. We took him up in the air and the only test we could do quickly was to go up and start ground-to-air communications and see how he made out. Well, I can tell you he did a lot better than I did, because he knew what he was expecting." The conclusion is quite

inescapable: expectation based on the situation, or the previous linguistic content, is a very strong factor influencing the perception of a speech signal. This factor is sufficiently strong to over-ride judgments which, if based on the syntax or phonemes, would differ from what the listener feels is a reasonable utterance in the present context. Bringing into play questions of reasonableness and appropriateness of the preceived response involves the listener's past and present experience. By virtue of this, the entire organism must be considered in the act of perception. However, such an explanation must be considered as a rephrasing of perceptual questions and not as an answer.

In linguistic terms, the essence of semantic content is represented in terms of the deep structure which is much to close to "thought" to be discussed in this thesis.

5.4 DIFFERENTIAL ABILITY OF SUBJECTS

Although the results of the sentence comprehension tests furnish conclusive evidence of the fact that transformed speech can be learned, the scores for these tests range from 100% to 5%, even on the last session. The extreme range of performance adds a dimension of complexity by its implication that the adaptation mechanisms are a function of the listener's cognitive organization, or in more general terms, a function of his "personality". Thus, in addition to involving speech and linguistic habits, comprehension of transformed speech requires that the subjects exhibit a plasticity in perception, that is, a learning of new strategies for using acoustic cues. This, however, brings into play questions of attention, motivation, mental blocks, etc., which are discussed in Section 6.2.

The consequence of large variations in scores is not severe since the statistical correlation of a subject's performance from session to session is very high for most tasks. That is, a subject who perceives sentences well on a given session is very likely to do so on the succeeding sessions. The observed consistency over the course of the experiment allows the relative abilities of the subjects on

different tasks to be compared. One might conjecture, for example, that if the perception of one-syllable words requires the same cognition as the comprehension of sentences then a subject who was particularly talented on one task would be talented on the other. This is the same arguement that was used in the correlation studies discussed in Section 4.6, and it will now be used to demonstrate the relation between the various comprehension tasks.

The scores for each subject on every test were converted to ranks, and the Kendal Correlation of Concordance was used to determine relative consistency of performance. A difficulty arises, however, from the fact that for many of the comprehension tests, especially during the first half of the experiment, the scores are so low that a rank measure is not very reliable. On the two-syllable word test of session one, for example, all subjects, with one exception, could not perceive any words correctly. Therefore, correlating the ranks on this test with those of the same test given on session eleven is not meaningful. The same situation occurs for other comprehension tests.

The ranks for the sentence comprehension tests given during the second half of the experiment are very highly correlated with $W = .88$ ($p<.001$); but when the tests for the first half are considered alone or in combination with the second half with $W = .33$ (not significant) and $W = .34$ ($p<.01$) respectively, the correlation is observed to be much weaker. The lack of a correlation among the test scores for sentences given in sessions one through seven is the result of very low average performance. In other words, a subject who perceived a few extra words correctly, or a subject who felt like guessing, would raise his score above the others and could change his rank from twelve to one. It was, in fact, found that in the beginning some subjects were much more willing to guess without being confident that they were right, while others were more conservative. Those who guessed generally scored, relatively, very high initially but their scores did not increase with more practice. Furthermore, the only words perceived correctly in the beginning were function words, as mentioned in Section 5.31. Later, the ranks are determined by the number of content

words comprehended since there are many more content words than function words.

The scores for the word tests all remained low throughout the experiment, and the ranks did not correlate with each other, although the geographical-names-test results were significantly consistent with $W = .51$ ($p<.01$). After examining these results it was noticed that the geographic-names-test scores for the first and last session had a very high Spearman correlation $R_s = .87$ ($p<.001$), but this set of ranks did not correlate with the other presentations of the same test on sessions three, six and twelve. This might be explained on the basis of identical test material, however, the material on session three was the same as on session twelve and the correlation was small, $R_s = .50$ (not significant). The essential difference between the two sets of tests is that there was no conversational practice on the first and last sessions, which suggests that these results were influenced by extinguishing of the perceptual ability or that the results were more dependent on the content of the tests than one would have hoped. In either case, there is no reliable evidence available from the experimental data to examine these questions.

In order to correlate performance for different tasks, the rank scores for every test of a particular kind are averaged and then re-ranked*. The most significant finding from the correlation studies is that the performance for sentences (second-half of experiment) and geographic names have a Spearman coefficient of .88 ($p<.001$). Such a high correlation, quite unambiguously, states that these tasks are using the same cognitive mechanisms. The perceptual cues in both cases can be seen to be very similar since stress pattern, syllable length, and semantic redundancy all play a role. The subject who was able to identify every name was also able to comprehend every sentence.

The tests using word categories also illustrate another point. When the subjects were given the names of the categories to which the

*See Section 4.6 for a more complete discussion of this technique.

words belonged, the scores correlated somewhat with the sentence tests (R_s = .63, $p<.05$); but when the category names were not given so that the subject had no semantic redundancy cues, there was no correlation with the sentence tests (R_s = -0.15). Since the stimulus words were the same perception based on the phonemic sequence would have shown a very high correlation. The lack of correlation must, therefore, be interpreted as meaning that semantic redundancy, rather than sensitivity to the spectral distortions produced by the system, is the determinimg factor in learning to perceive. This would suggest that the differential ability of comprehension is a manifestation of a subject's ability to use non-acoustic information to compensate for a lack of distinct phonemic preception.

5.5 RELATIONSHIP BETWEEN COMPREHENSION AND PHONEME PERFORMANCE

The suggestion in the previous sections that phonemic performance is not a good indicator of sentence comprehension is further borne out in the correlation studies to be discussed. Although sentences are composed of a sequence of words which are composed of phonemes, comprehension of a sentence does not result from identifying the word or phoneme elements. Because of the complex structure of a sentence, the listener can, and probably does, use the syntactic and semantic redundancy to aid in perception. It might be stated that words are more related to a sequence of phonemes than sentence are related to a sequence of words. Sentences have a unifying prosodic structure which appears to act as a separate channel for carrying linguistic information. The exact nature of prosodic information cannot be determined from this experiment but it is, nonetheless, a very significant aspect of the perceptual problem.

Of the four kinds of phonemic tests measures discussed in Section 4.6, consonant identification, vowels identification, phoneme discrimination, and language discrimination, only the first is found to be related to a subject's ability to identify isolated, contextless

words. The Spearman rank correlation* for word identification and consonant identification is .73 ($p<.01$). There was no significant correlation with any of the other tests. Since consonants carry the bulk of linguistic information, and since there were no other cues, this result is hardly surprising. In a one-syllable /C V C/ word, the listener cannot use stress, syllable length or context. He must guess from the set of words which are in the same form, and his probability of getting it correct is directly dependent on his ability to perceive the phonemes. The lack of any correlation between word perception and vowel identification is most likely another manifestation of the vowel instability phenomenon, discussed in Section 4.24. Whereas, the consonant errors in word perception are primarily in the place of articulation feature, the vowel choice appears to be unrelated to the actual stimulus vowel. It is impossible to state if this is also true for untransformed natural speech or if it is particular to this experiment.

In contrast to the perception of isolated words, the comprehension of sentences and geographic names is not correlated with either the phonemic tests or the word tests. In other words, a subject who can do better at identifying or discriminating a transformed phoneme may very likely not be able to comprehend a sentence. Equally, a subject who cannot perceive phonemes and words in isolation may be able to comprehend sentences. This must be interpreted as meaning that when the stimulus is as complicated as a sentence, or involves the same cognitive processing used for sentences, the listener uses a different kind of strategy than he does in identifying discrete acoustic units. This conclusion would give support to those theories which maintain that the units of perception must be as large as syllables or words (Miller 1962). Moreover, if perception were a hierarchy of operations

*In performing the correlations, the technique described in Sections 4.6 and 5.5 is used. First the average ranks for a particular set of tests is found and then re-ranked. Then, these ranks are combined with the rank scores for similar kinds of tests. For example, the ranks for all the one-syllable word tests are re-ranked and then combined with the re-ranked scores for the two-syllable tests to form and overall measure of relative word identification ability.

then one would expect that sentence comprehension would be dependent on word performance, and word performance would be dependent on phoneme performance. In fact, this does not seem to be the case. An alternative theory would maintain that cognitive processing is a parallel operation and that the prosodic features of a sentence can act to supplement word perception in a sentence context. The lack of correlation between the different phoneme tasks and comprehension rules out the possibility of any simple theories.

CHAPTER VI

LEARNING SPECTRALLY TRANSFORMED SPEECH

The results of the testing series, although objective and structured, are controlled measurements and, by virtue of this fact, do not represent many facets of speech perception that would occur in a natural environment. Even the sentence comprehension tests, which attempt to measure comprehension by scoring the percentage of syllables correctly identified, are not based on an attempt to communicate an idea. In contrast, a normal listener engaged in conversation tries to understand the meaning of the speaker, more often than not, without actually remembering the words which conveyed it (excluding phoneticians and psychiatrists). This is illustrated by the fact that when a listener is asked to repeat what he has heard he generally paraphrases the thought rather than repeating it word for word. People with good memories can often recall a sentence verbatim without actually understanding the thought contained within it. Thus, the relationship between perceiving and understanding may be as different as phonemes and sentences. The listener engaged in conversation is attending to the most abstract level of cognition--the idea.

6.1 PSYCHOLOGICAL CONSIDERATIONS OF CONVERSATION

A conversation is a dialog with each person alternating as speaker and listener. The speaker, for this reason, receives feedback information about how well he is being understood and, in addition to correcting misunderstandings and ambiguities, he can control the rate with which new information should be introduced. The content of a sentence is somewhat redundant and predictable since it is related to the previous sentence. The speaker controls the semantic redundancy, for example, by repeating phrases and clauses instead of using pronouns. Of all the tests, the content-related sentence test is probably most similar to conversation.

Conversational utterances differ from the test sentences in

other ways. Typical utterances are generally fractured, being only parts of sentences, and are sometimes ungrammatical. However, conversational speech is often richer in prosodic features since the speaker is conveying his attitude, such as displeasure, sarcasm, doubt, surprise, etc., about the content of the utterance. Also, many standard phrases are used and the vocabulary can be selected to avoid confusions.

The undefined and ambiguous nature of conversational speech makes it very difficult to analyze, especially since the subjects were in the process of learning as well as communicating. Although perception, in general, is an active process, the testing situation was passive in comparison to the conversation situation. Because of its dynamic aspect, and because many different kinds of psychological activities were occurring at the same time, the 42 hours of tape recorded conversation data can only be analyzed by first establishing a set of hypotheses. This, in turn, makes the insights gained from conversation subjective and open to question. Nonetheless, the experience of the subjects during the conversations is a very important factor reflecting their ability to learn transformed speech and must not be ignored.

Many of the practice techniques used by the subjects and many of the learning stages through which they passed were analogous to other language situations. In trying to understand a speaker with a foreign accent or speech defect, for example, the listener is faced with an utterance composed of the same lexical items and the same syntactical grammar of his own language, but the pronounciation of the phoneme sequence and the prosodic features are distorted. Similarly, the infant learning to speak is at first exploring a new sensory medium and is correlating his production of a sound with the perception of it. Also, learning to understand a foreign language requires that the student integrate new sound units with already existing symbols. The kinds of exercises used by language teachers are somewhat similar to the activities of some of the subjects.

The experiment with transformed speech began by immersing the

pair of subjects into the new medium, without visual contact, preceded with the instructions "you must learn to communicate with each other and that is your only goal." The advantages of using an unstructured paradigm are twofold. Firstly, there is no conclusive experimental or theoretical evidence to indicate the nature of a learning program which would maximize learning transformed speech; consequently, a pre-defined learning algorithm would make the results of the experiment very dependent on the effectiveness of the designed practive sessions. Moreover, it is difficult to create an algorithm which takes into account the different kinds and rates of learning for each subject. Secondly, it was felt that insight could be gained by examining the relationship between the types of practice and results on comprehension tests. Since there may be more than one cognitive mechanism involved in this learning, it is hoped that the stages of learning reflect stages of practive activity. By not biasing the subjects with specific tasks or instruction, they were effectively being told that there was no "right" approach. They definitely had the feeling that they were pioneers exploring a medium which nobody, including the author, understood.

The subject pairs, upon first being placed in the spectrally transformed medium, started talking as if they were in a normal acoustic environment. Examples of their first utterances are "Hello Sam", "How are you?", "Some son-of-a-bitch stole my lab notebook.", and "George, can you hear me?". Hearing their own and their partner's voice sounding so strange provoked amusement and then the profound realization that they were not being understood. At this point learning began.

6.2 ROLE OF PERSONALITY

A subject's response to the realization that he was not able to communicate in transformed speech medium appeared to be a function of his personality and, for this reason, there were a great variety of reactions to the situation. After the first several seconds of conversation attempts, several subject pairs asked, almost simultaneously, "do you understand?". Since they both had the word "understand" in mind, this became the first word which was communicated. Similarly,

the words "yes" and "no" were discovered in conjunction with the many repetitions of "understand", but any utterance more complicated than these words was met with the response "do not understand" or "understand no". In contrast to these pairs, one set of subjects continued talking as if they were being understood; each one was talking about a different topic and was assuming that his partner's responses were related to what he said. These subjects, by not listening carefully to their partners were able to pretend that everything was in order. Denial was their way of coping with a situation which they felt unable to handle.

It becomes apparent when listening to the tape recordings of the conversations that the subject's "self-image", his attitude to a problem situation, and other emotional feelings are, in addition to his linguistic ability, a very strong factor in determining his approach. Overcoming the lack of intelligibility produced by the spectral transformation required the subjects to generate a strategy for learning and practice. Thus, their ingenuity, as well as their perseverence, played a role. They were continuously faced with the questions: "how successful is our present method?", "how do we measure our success?", "have we done as well as we ever will?", "should we keep trying new approaches until we find one which works?", "should be persist with the method that we feel should work?", and "can we ever learn to understand each other?". The answer arrived at, either implicitly or explicitly, are a function of their personalities.

The effect of personality differences were illustrated in many ways. A person who, for example, liked an unambiguous situation would work best in a "directive" as opposed to a "non-directive" environment and would, therefore, find the unstructured instructions of this experiment frustrating. Whereas some of the subjects often asked the author "what should we be doing?" and "how well are we doing compared to the other subjects?", others liked the independence and worked at exploring the medium, ignoring their partner. The mere fact of having a partner raises the issue of cooperating in a team effort. Internal competition between the members of a pair might act to stifle tendencies to

explore and to make mistakes. Since the subjects were not rewarded monetarily for learning to communicate, the motivation to exert the mental energy necessary to learn was internal and personal. Just as some students learn better in a classroom for the reward of grades, those subjects who did not learn to understand transform speech because they did not like the freedom or lack of a grade might have performed better if they had been directed to engage in specific activities and exercises with a numerical measure of their success.

Can, it must be asked, the wide range of performance be attributed to the personality differences? For example, should the lack of ability to communicate be interpreted as the basis for becoming bored and losing interest in the experiment, or should a lack of interest and ingenuity be interpreted as the basis for a failure to learn. A subject who tries to master the situation with only one approach that does not succeed might well give up, although if he had used a more effective technique he might have succeeded. In conclusion, the nature of the conversational practice sessions determined the learning experience and, as will be shown in Section 6.5, is directly related to test performance. The relative amount of time spent in different kinds of activities correlated with the rank scores on many of the tests.

Although the partner of each subject was his friend, one of the two generally had a stronger interest in the experiment, or a stronger motivation to take control of an unstructured situation, and would dominate his partner. The dominant partner, by forcing the submissive partner to act as responder, was more influential in determining the kind of practice that each was experiencing. For the pairs who were able to co-operate and were equally motivated this was not a problem, but for some the imbalance was apparent. If one subject, then, did most of the talking, the listener was gaining experience in perception, assuming he was paying attention, while the speaker was not. Under this condition, which sometimes approached an extreme, one subject was experiencing a very different kind of practice than the other one. No attempt was made to interfere with their working relationship even

though it might not have been optimum or even useful; however, they were reminded that their goal was to learn to communicate and that this was their only goal.

Based on purely subjective insight, there appeared to be three personality types. The first type had a highly analytic orientation and approached the problem of communication by trying to find the rules for spectral rotation. This kind of personality, characteristic of many M.I.T. students, approached the situation initially by organizing and ordering the sound units. The second type appeared to have less interest in technique and merely tried to converse hoping that communications would suddenly happen. The range of strategies was limited and they seemed to have very little of interest to say to each other. Their attitude was relaxed, although also passive. The third and most successful type used a systematic approach, in that they experimented with techniques, measuring their success carefully, but they concentrated on the communication problem as a transfer of ideas. They differed from the first type in that they were not interested, per se, in spectral transformation.

The importance of personality in relationship to linguistic ability has been shown in many foreign language learning studies. Anisfeld and Lambert (1961), for example, found that the ability to learn Hebrew correlated with the students attitude towards the Jewish people. Those students who were willing to identify with the social group speaking the language found it easier to learn the language. Lambert (1963) stated "it was clear that students with an integrative orientation were more successful in language learning in contrast to those instrumentally oriented. Further evidence indicated that this integrative motive was the converse of an authoritarian idealogical syndrome, opening the possibility that basic personality dispositions may be involved in language efficiency".

Moreover, students of a foreign language show differential ability to learn different language tasks, as illustrated by the fact that

a visually oriented person learns to read the language better than he understands it aurally. For aural comprehension, Karlin isolated two factors which he considers relevant: the ability to fuse perceived parts of an auditory stimulus into an integrated whole and the ability to remember a large auditory span (Blickenstoff 1963). Other studies have opened the possibility that language ability is related to musical or pitch acuity, although they are by no means conclusive or consistent. Even though the factors which contribute to language aptitude are not known, it can be safely said that personality differences, as reflected in different cognitive strategies, allow some people to learn a new foreign language with ease and to speak any of a half-dozen languages, while others never learn a foreign language fluently. In this sense, learning to perceive spectrally transformed speech is probably related to many of the cognitive elements which play a role in foreign language learning. No attempt was made to measure the language aptitude of the subjects in the experiment.

In addition to the personality and emotional factors, the speech habits of the subjects were important. The clarity of a speaker's articulation, even in normal speech, determines how readily he is understood especially in the presence of background noise or distortion. While some of the subjects had a strong tendency to mumble, assuming that they were not going to be understood anyhow or assuming that their speech pattern was not important, the subject pairs who learned to communicate most quickly were the ones who were very careful about their enunciation. A clear observation of this phenomenon was shown when "biology laboratory" replaced "bio-lab". Also, talking too rapidly or too slowly decreases intelligibility since rapid speech is often slurred with many sounds being mis-articulated or omitted and, excessively slow speech has an unnatural stress and intonation pattern with many phonemes inappropriately lengthened. Furthermore, when a sentence is spoken too slowly, it cannot be remembered, and therefore perceived, as a unit.

The difficulty with which an idea was communicated depended on

its complexity and familiarity. Thus, two roommates who had often discussed the subject of women had less difficulty in understanding sentences relating to this theme since they already knew what each thought about the subject. Even though the conversations were not rehearsed, familiarity with the idea being expressed increased its redundancy and increased its intelligibility. Similarly, conversations about current events were more intelligible if both partners had heard the newscast or read the newspaper. Aside from content, the length of time that the subjects had known each other was a measure of how accustomed they were to each others syntax, idoms, manner of speaking, etc. The pair of twins in the pilot study learned to communicate most rapidly and their initial conversations were composed solely of frequently used personal idiomatic expressions.

The above discussions point out the complexities of the experiment but they do not help, nor are they intended, to evaluate the influence of personality on learning. Questions of personality and learning are much less understood than questions of perception; and questions of perception are not well understood. Since the goal of the experiment was to demonstrate that spectral pattern transformation does not prevent communications, the fact that some subjects did learn to comprehend transformed speech is sufficient. In the same way, one does not have to show that all people can learn a foreign language in order to prove that it can be learned.

6.3 QUESTIONS OF LEARNING

In designing a learning experiment, one must specify two parameters: the length and frequency of the practice sessions. Assuming a fixed amount of available practice time, it should be distributed in such a way as to maximize learning. As shown by the discussions in Woodworth and Schlosberg (1961, cht. 25) there are many factors which influence the efficiency of learning.

Towards the end of a long learning session, fatigue and boredom, called "reaction inhibition", reduced the learning efficiency to the

point where further practice resulted in no improvement in performance and a possible deterioration in performance. The saturation point occurred when the subjects were no longer able or willing to expend the mental energy necessary to try to perceive the spectral transformed speech. The amount of conversation time required to reach saturation varied as a function of the kind of practice and the attitude during that particular session. In the beginning, when the subjects were very enthusiastic, they were able to maintain active interest for the full half-hour. Those subjects who later found conversation possible, and if they also had something to say, were also able to maintain active interest for the full session. When the session was not going well because they were not able to communicate or were dissatisfied with their level of performance, they would lose interest after about 20 minutes.

In most learning situations, a certain amount of forgetting or unlearning takes place during the interval between the sessions. Thus, in addition to the normal warm-up time required to get into "the swing of things", a certain amount of time at the beginning of each session is expended on raising performance to the level of the last session. The forgetting that takes palce is, however, somewhat more complicated since it is often observed that performance improves during this time. The subjects, during the beginning of the experiment, sometimes became fixated on a given sentence or a particular approach in such a way that the time spent did not appear to enhance performance; but, the next session allowed a fresh start without the residual feeling associated with an unsuccessful attempt to communicate.

Increasing the interval between learning sessions, in some experiments, actually reduces the total amount of practice time needed to master a task. Bumstead (1940; 1943) showed that the total amount of time need to memorize prose or poetry decreased when the practice time was distributed with larger intervals between sessions. The total amount of time required to reach a given level of performance was approximately the same when the intervals ranged from two to eight days. It has been argued that in this kind of task the time lapse between sessions requires that the learner utilize those associations which are

not perfectly fresh and ready but have become habit. The learning of spectrally transformed speech may not be directly analogous to memorizing so that it is difficult to apply Bumstead's results to predict the optimum interval between sessions in this experiment. If learning the transformation is related to long- and short-time memory, then the suggestion of Bilodeau and Levy (1964) that there is a similarity between verbal and motor skills and retention, and that there is a large drop in retention after a few minutes but a very slow drop during the next week, could be applied to learning transformed speech. The difficulty in assessing the situation is caused by our ignorance of what perceptual learning is.

When the twins of the pilot study were tested about a month after their last learning session they showed a very high degree of retention. Only a minimal amount of practice was required to bring them up to their previous level of performance.

6.4 STAGES OF LEARNING

The content of the conversation sessions was analyzed by classifying the techniques used into one of four categories or stages. Since the subjects progressed, on the average, from one category to the next, it was felt that the categories are somewhat related to stages of learning. There are, however, significant exception and variations. Some subjects skip a stage entirely, others spent a disproportionate amount of time in one kind of activity, and others jump back and forth between different strategies. Nonetheless, the scheme used in the following discussions gives some insight into the processes involved in mastering transformed speech.

6.41 STAGE I ACOUSTIC PROBING

Most of the subjects spent part, or all, of the first session exploring the properties of the acoustic medium. Often, they would utter a continuous diphthong-like vowel, as in "aaaaaeeeeooooo" or a sequence of simple monosyllabic phonemic segments, while listening very carefully to the transformed sounds in their headphones. When operating

in this mode, they were ignoring their partner and concentrating on their own perceptual-motor activity. In one sense, this kind of acoustic probing might be viewed as learning to correlate the perceived sounds with the articulatory movements which gave rise to them. Fry's (1966) description of infant babbling shows some marked similarities to this kind of acoustic practice. "Utterances are characterized by frequent repetitions of the same syllable or sound, and the significant feature of this stage is that the child is now uttering sounds for the pleasure they give him and not as an expression of his reactions to some particular situation." When first exposed to transformed speech, almost everybody tries to make as many "funny" sounds as they can simply to enjoy the exotic and unnatural nature of spectrally distorted sounds. Whistling, singing, and periodic pitch fluctuations all provoke amusement. For the babbling infant this stage is characterized by the establishment of the auditory feedback loop, in that, a strong link is made between the auditory perception and kinesthetic sensation. However, with transformed speech the situation is somewhat ambiguous. Because the listener can hear his own voice through headbone conduction, he can consciously choose to listen to either the transformed sounds or the untransformed sounds. Thus, it is impossible to say if he is correlating new sounds with motor-movements or with old sounds. Nonetheless, this stage is exploratory.

Some subjects used this initial period to enhance their discriminability of phonemes by practicing sequences such as "mat, pat, dat ...". The inability to quickly classify transformed sounds into normal phonetic categories, as well as the inability to immitate them, makes them difficult to distinguish. The student of a foreign language is also faced with having to make distinction in the new language that do not exist in the old, as for example a Japanese learning the /l/ and /r/ differences. Calvert (1963) and others state that in intensive Peace Corps Language training students spend time practicing pronounciation and discrimination of non-native sounds. Weir (1962) observed that her child, after babbling stage, practiced rhyme sequences not too dissimilar to those of these subjects.

Very soon after beginning acoustic probing, subjects started to use other stimuli, such as the alphabet, counting, and nursery rhymes like "Mary had a little lamb". Both partners were now generally active; after listening to an utterance, one subject would try repeating it with his partner now acting as listener. Although there was sometimes a diffulty in establishing what was being said, once the listener knew that his partner was counting, for example, he would have no trouble understanding the sequence "one, two, three....". Other examples of this kind of practice included reading from a text which both subjects had in front of them. One pair of subjects brought in a list of numbers followed by topic words. They would say the number followed by the word; since the number was understood, they would listen carefully to word associated with it.

This kind of practice must be interpreted as occurring at the acoustic and phonemic level since they are practicing with specific sounds rather than with word groups containing ideas. There was never a question of conveying information because they always knew what was being said. As evidence that they were being understood a subject would repeat the word, and his partner would repeat it again. Their attention is being focused on the perceived sounds, and it may be speculated that the cognitive process is that of remembering sounds or correlating them to their word symbols.

The most natural extension of phonetic practice was learning to imitate the transformed phonemes. A few of the subjects, whose technical background told them that an accurate pronounciation of a transformed sound should appear normal to his partner after begin transformed again by the system, tried to learn to speak transformed speech. Even those who did not know the above fact tried to pronouce the distorted sounds. In effect, they were attempting to ascertain if it would be easier to speak in an unnatural way than to perceive unnatural sounds. Two pairs spent several sessions in this kind of practice. They happen to be particularly adept at modifying their pronounciation so as to find the correct pronounciation for particular sounds in isolation. For

several reasons their success was only moderate. Firstly, many of the consonants cannot be imitated simply because there is no articulatory movement which duplicates the inverted formant structure. Secondly, they were not aware that a consonant is not defined in isolation and that it strongly dependent on the neighboring vowels. Thirdly, the perception of the vowels, as discussed in Section 4.24, is unstable; that is, the perception is not directly related to the stimulus. Fourthly, a spectrally transformed word is not a sequence of transformed phonemes. Fifthly, they found it difficult to remember the pronounciation of a word they had learned to pronounce correctly.

One pair had mastered the transform of the word "English" and many of the letters of the alphabet. They discovered the consistent transform pair /u/ and /i/, also /y/ and /w/, which are the most clear-cut illustrations of front-back reversal. They used this learned ability mostly in spelling a mis-understood word. Later, during the middle of the experiment they attempted speaking transform speech mainly as aid to normal comprehension, but it would often take many attempts to find the right pronounciation of a given word. Eventually they abandoned the approach because they were becoming confused as to when their partner was speaking transformed and when he was speaking normal English, and because their ability to successfully imitate many words was poor.

Their experience with production, however, does illustrate some rather well known facts. Speech production is a more active process than comprehension and is often thought to be "ballistic". Speaking requires that articulators be enervated with a highly synchronized set of signals incorporating past, present and future information. The entire word sequence is prepared in advance and delivered automatically, without feedback, to the production apparatus. Production, being a motor activity, is characterized by a sequence of irrevokable decisions. In contrast, perception with the aid of short-term memory allows for processing, differentiation, and integration while additional verbal information is being accumulated to resolve ambiguities. Ehrman (1963) makes the well known point that "listening for comprehension and oral

expression are two different skills which are not acquired at the same time with the same ease and speed when learning a foreign language. Our ability to listen to a foreign language and to understnad it is greater than our ability to speak it."

In clinical experiments, Ringel and Steer (1963) showed that articulation is relatively stable even in the absence of feedback. Articulation error scores under the conditions of either high masking noise to block acoustic feedback or nerve blocking anesthetic to remove kinesthetic feedback increased only slightly. With both kinds of feedback removed, articulation errors increased significantly, but the speech was still quite intelligible. Thus, speech production appears to be organized in a feed-forward manner and does not require extensive feedback. Throughout the experiment subjects were listening to their own voice transformed, yet their articulation remained unaffected. If feedback were necessary for good articulation, one would expect that subjects would begin to say /we/ for /you/ or /oat/ for /eat/. They did not, however.

6.42 STAGE II: HIGH-REDUNDANCY UTTERANCES

This stage is characterized by conversations which, by virtue of their high predictability and limited range, contained very little information. In contrast to Stage I which was fragmentary and phonemic, Stage II conversations were structured but simple. There were basically two types of strategies included in this group: limited-set encoding and restricted topics. In the former, subjects tried spelling words that were not understood or using a number code to correspond to a particular topic. In the latter, the content of the conversations was limited to words and structures which were highly associated or parallel in form. In many ways, the character of this stage is very similar to the drill practice given language students. Although the cognitive energy appeared to be focused on perception of content words, the subjects did not communicate large ideas. Rather, they measured their success on the basis of identifying a key word or phrase.

At various times throughout the experiment, most subjects used the techniques of spelling words that were not understood. Only a limited amount of practice was necessary in order to identify a letter. The technique of learning the letter was to recite the alphabet up to and including the letter to be transmitted. Thus, for example, if the word "bad" was not understood, the speaker would say "ab...a....abcd". Spelling represents a code of only 26 elements and, therefore, has a much higher intelligibility than words. Moreover, the 26 elements can be perceived in a sequence of independent non-interacting linguistic units. Even though this method was effective, its inefficiency was apparent to the subjects and they only used it when the specific word was not understood. Often spelling the important content words and then repeating the sentence in which they were used resulted in total comprehension.

Numbers are a linguistic set analogous to letters in the alphabet. Very ingeniously, one pair of subjects on the second session brought with them an index, as shown below:

1- house
2- weather
3- motorcycles
4- summer
5- humanities
6- talk in French
7- un homme et une femme
8- yes
9- no
10- repeat
999- panic

This shows that these subjects understood that the difficulty in expressing independent words and ideas could be avoided by using a simple coding scheme. At first they used their index to indicate which word was being spoken, but they then realized that they could converse about a subject that was identified with the number. They initially went through the last several times learning to correlate the numbers with the word names. Following this practice they went through the following sequence:

a) "two weather, two weather"
b) "weather"
a) "right"
b) "weather rain"
a) "raining"
b) "weather wet"
a) "the weather is raining"

Although the conversation was actually more labored than it appears above, they were using the technique of categorizing and limiting semantic content to enhance perception. This pair of subjects also demonstrated that they could distinguish French from English. At one point, one subject began counting in French, "un, deux, trois.....", and his partner responded, asking him if that was French. The question of transfer of learning from one language to another is not answered in this experiment.

Another pair of subjects read to each other from a newspaper. In this way they knew what was being said and listened carefully to the way it sounded. As a check on their ability to understand, they selected sentences at random, after they had read it through, trying to identify which sentence was spoken. This was very successful owing to the limited possibilities. Towards the end of this session one subject asked the other "Are you going to eat at Walker". The word that was perceived was "Walker", following which the listener said "I am going to eat at Walker". The next comment was a synthesis of the name Walker and an article read during the earlier part of the session. He said "Did you know about the kids who got food poisoning at Walker", which was understood. Again, it is found that comprehension was based on categorization and familiarity. But, in addition, they discovered that names were particularly easy to understand.

The use of names of cities, states and countries as practice material was tried by one pair of subjects with large success. From this they moved on to the names of books, also with great success. Other subjects discovered sentence groups such as "There are 365 days in the year", "There are 12 months in the year", "There are 4 weeks in the month", etc. Once focused on this kind of theme, they were able to

communicate the names of the months and days without difficulty. In general, most subject pairs developed a frequently used ritual vocabulary including such phrases as "repeat", "say again", "very good", "you are stupid", "talk faster", "talk slower", etc. Or, they would interject common phrases such as "peter-piper picked.....", "the rain in Spain stays mainly on the plain". Conjugating verbs was also practiced.

The subjects were focusing on content words and sometimes they would put a correctly perceived word into the wrong sentence or turn a statement into a question.

From the results of the sentence tests, it is quite apparent that the subjects could have known, if they wanted to, that they had not correctly perceived the syntax and meaning of the sentence, but they allowed their perception of a content word to dominate. For example, if the speaker said "I am going home for vacation", the listener might repeat "you said, what am I doing this vacation". This phenomenon might appear paradoxical in that the structure of the stimulus sentence and the perceived sentence bear no relationship to each other, aside from the word "vacation". This must be interpreted as meaning that the subjects, at least in the intermediate stages of learning, concentrated solely on the important words and not on the overall patterns. In a sense this stage reflects an attempt at perceiving word symbols and not ideas and thoughts.

For most of the subjects this practive was rather inefficient. They would often spend most of the half-hour session communicating only a few words and ideas; the remainder of the time was spent trying, in many ways, to communicate a given word. Frequent repetitions, generally, did not help beyond the first few. Probably a carefully designed initial learning program would have increased the rate of learning, although perhaps not. This appears to be the critical learning stage, since in Stage III subjects could communicate a given idea using semantic strategies.

6.43 STAGE III SYNTHETIC CONVERSATION

This stage, in contrast to Stage II, is characterized by communications of ideas and information in a dialog approaching conversation; but it is not true conversation since the subjects still had a great difficulty with many important words. The technique used to overcome a non—comprehended or misunderstood word consisted of semantic branching or ideas by association. The conversations were somewhat similar to those which occur in the Peace Corps language training centers where students are required to speak the new language when socializing at dinner. For example, if "pass the salt" was not understood, it might be followed "pass the thing next to the pepper" or "give me the spice to put on my meat". The listener can intuit what the speaker means by associating content words. "Salt" and "pepper" are often heard in the same context.

From a psychological point of view, it is during this stage that subjects experience a rather sudden improvement in communications, called "break-through". They become much less cautious about the kinds of utterance they will attempt and they begin working on the assumption that they will be understood. Generally, any attempt to practice learning is abandoned; rather, they simply try conversing with the expectation that they will eventually be understood. It is almost as if the subjects "lock" on to the speech patterns and no longer listen to specific words.

The following dialog illustrates the use of content related words for communications and is the kind of conversation which occur just before "break-through".

C So how was the folk dancing?
D Today is?
C the folk dancing
D you said today?
C What kind of folk dancing?
D I don't get it

C Was it American folk dancing?

D American?

C yeh! American folk dancing

D American state Johnson

C dancing!

D What is the last word?

C dancing. What you did last night

D Oh! folk dancing

C folk dancing, right

D Did I like it?

C it was American?

D Oh yeh! I like it. Did you say American Folk dancing?

C Was it American?

D No, mainly, mainly Balkan.

C What kind?

D Balkan, Russian, Slavic, Hungarian, Rumanian.

C Ireland?

D Hungarian, Rumanian, understand Hungarian

C I do not think so.

D from Hungary

C India?

D from Hungary

C India?

D Rumania, Hungary, Russia, the Union of Soviet Socialists Republics

C dancing?

D Russia, the biggest country in the world

C where is it? is it in North America

D No, it's in Asia and Europe

C where?

D Asia

C and Europe

D and Europe

C Russia

D the Union of Soviet Socialists Republics

C Russian

D Russia, right. Russia and Hungary

C what is the second one?

D Hungarian

C Where is it?

D Europe

C Europe

D Eastern Europe

C Eastern Europe

D Hungarian

C What is it again?

D Hungary, Poland, Rumania, Lithuania

C Is it part of the USSR?

D that's one, but also Hungary.

C Hungary

D Hungary, yeh

etc.

This example illustrates many points. The initial theme of the conversation took several rounds of interaction to be established and the first key word to be communicated was "American", a name of a country. The conversation proceeded to the point where "Balkans" was not understood; semantic branching then began. The names of the Balkan countries were given, but not understood. "Russia", "the Union of Soviet Socialists Republics", "the largest country in the world" were tried. The last was understood, although the dialog then had to locate the place with the words "Europe and Asia". After having communicated "Hungary", the conversation continued for another ten minutes in the same direction trying to establish the words "Balkans". To do this, they covered "Israel", "Jewish", "Palestine", "Jerusalem", "the Bible", "the country had a war", "the war in Viet Nam", "no Israel" and finally the idea was communicated. It is clear that the efficiency was not very great, but had the subject been able to guess the possible kinds of East European folk dancing, he would have perceived the word

"Balkans" much sooner. Nonetheless, the strategy is completely by semantic association, relating similar words in the hope that the correct word will occur to his partner.

The dialog also demonstrates the ease with which they used their verification and feed-back vocabulary. Such phrases as "where is it?" "what", "what kind", etc. were understood immediately. Repetition was used to verify correct perception but was not used to enhance initial perception. Some subjects found that only semantic branching, that is, using a new word instead of the word not understood, allowed an idea to be communicated; whereas others found that simple repetitions of part or all of a sentence resulted in comprehension.

The dialog is a clear example of inefficient communications of an idea using semantic branching. Although the word "Balkans" could easily have been communicated by spelling it, this pair of subjects felt that they wanted to persist with a direct approach. The fact that one idea took fifteen minutes to communicate does not mean that the time was wasted since they communicated many other ideas in the process of trying to establish a single word. Throughout this and similar dialogues, the emphasis is on thoughts and concepts rather than isolated sounds. The perceptual energy is directed at meaning, which contrasts with Stages I and II.

6.44 STAGE IV INTEGRATED CONVERSATION

The dialogues characteristic of Stage IV were simply normal integrated conversations with a minimal amount of repetitions, semantic branching, and other techniques. The content was usually general, not abstract or intellectual, and was often related to some specific experience or event. Only one pair of subjects in the main experiment and the twins in the pilot study really attained this stage for long durations. Other pairs were able to hold segments of conversations lasting about ten minutes but they had difficulties in changing the topic under discussion. Realizing that careful articulation was a very important factor to intelligibility, the pair who had the most successful

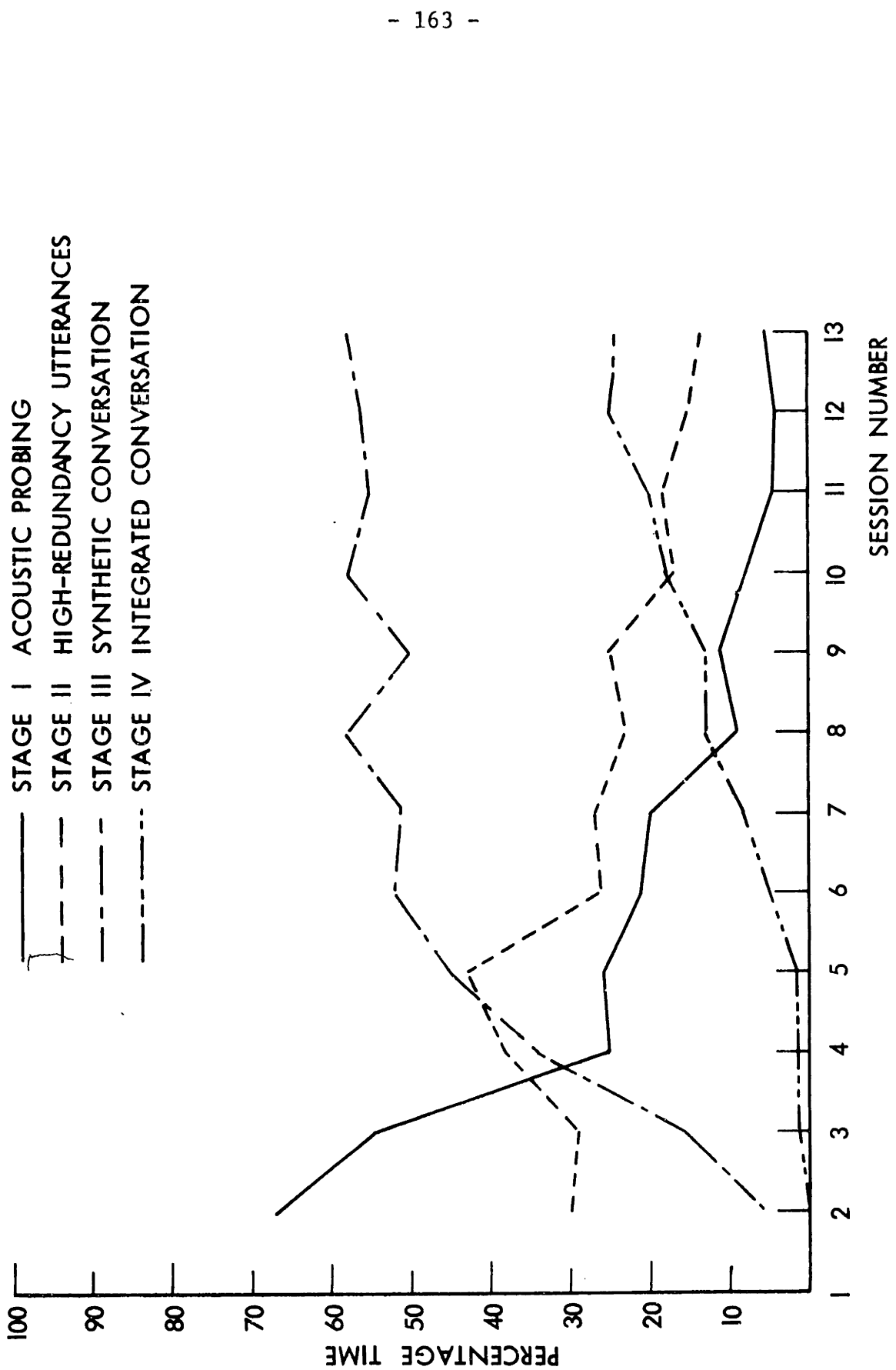
conversations behaved as if they were talking over a noisy telephone circuit. Nonetheless, they commented "this is just like sitting at home in the living room chatting".

It is perhaps interesting to note that there appeared to be nothing special about the intelligence, personality, or apparent verbal talent of these four subjects. If anything can be said about them it is that they appear to be "easy-going" personalities and approached the experiment as "fun" rather than as a challenge. Perhaps a more important factor is that the members of both pairs knew each other very well, being thoroughly familiar with each others speech habits and ways of thinking. In one case, the twins had known each other all their lives; and in the other case, they were roommates who shared many of their social activities.

6.45 CONCLUDING REMARKS

Although the stages of learning have been defined, illustrated and explained in the previous sections, it should be apparent that it was extremely difficult, and somewhat arbitrary, to classify a particular dialog between two subjects as being one kind of category. The technique used was to indicate the predominant type of activity for each 30 seconds of conversation for all of the practice sessions. Even with such small increments of time, there were often ambiguities and overlap caused by content which could, in fact, be classified as two types. For example, spelling a difficult word in the middle of a synthetic conversation is both Stage II and III. Because of this multiply defined dialogue, as well as total silence and complete non-comprehension, the actual figures used to derive Graph 6.4a should be considered to be only approximate.

The average data, shown in Graph 6.4a, demonstrates the validity of calling the different types of conversation categories learning stages. This Graph shows that acoustic probing was the dominant practice mode in the initial sessions, but that it was rapidly modified to include high-redundancy utterances. Real communications, however, did



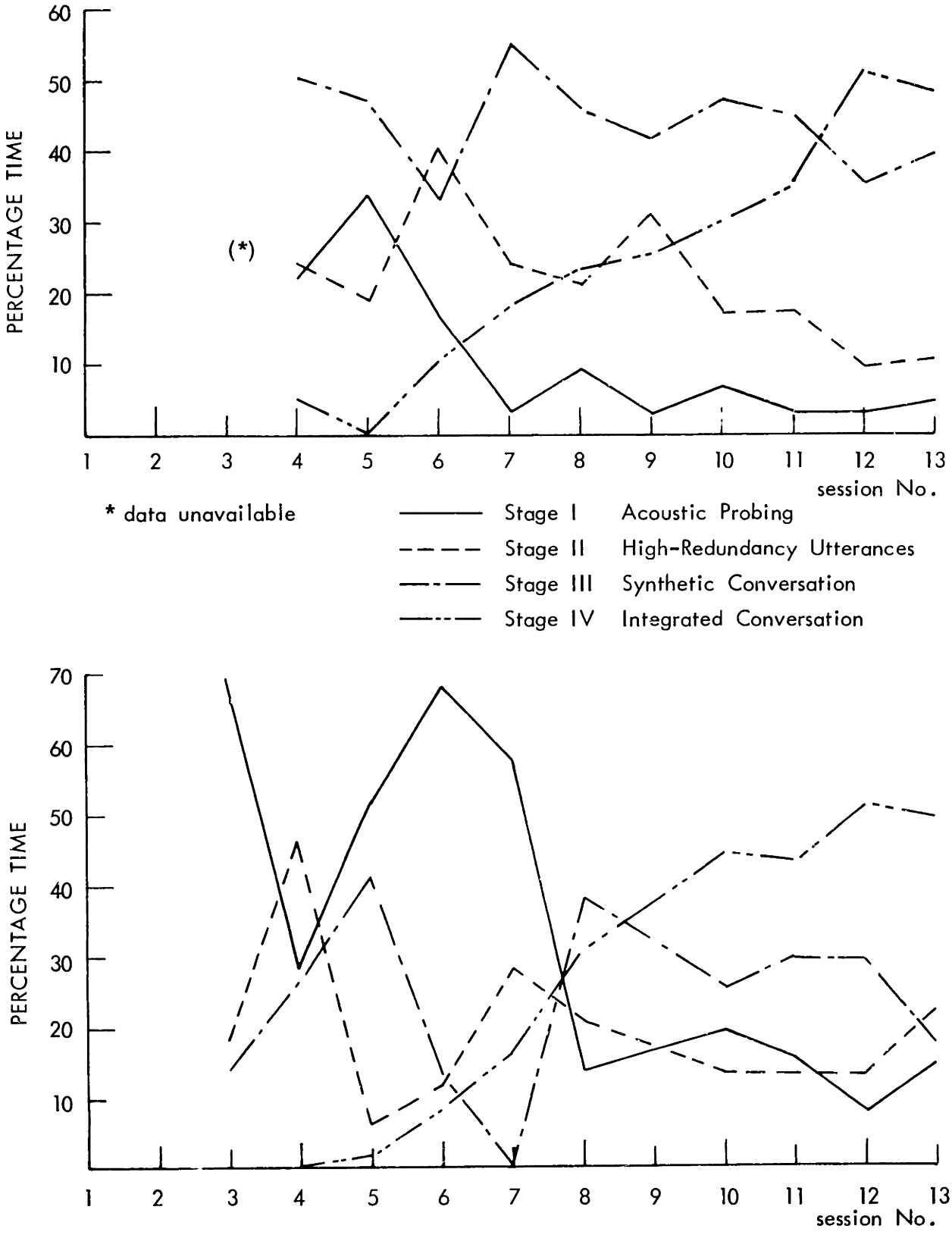
GRAPH 6.4a AVERAGE PERCENTAGE OF TIME SPENT IN EACH LEARNING STAGE.

not begin until later in Stage III. These conversations were very labored with only a few complete ideas being transmitted in one half-hour session. Natural conversation of Stage IV increased slowly and was the predominant activity at the end of the experiment for only a few subject pairs. Had the experiment been continued, or had the practive sessions been more frequent, more subject pairs might have attained the competence for free conversation.

The average data, rather than individual activities, of the subjects is indicated in the figure since any particular pair fluctuated in their kinds of practice. During a single session, a certain amount of time was spent in as many as four stages. The average scores, then, show a general drift from session to session. Moreover, if one of the subjects was tired, or pre-occupied, the subject pair would involve themselves in a less strenuous activity, such as phonemic practice instead of synthetic conversation, and this would manifest itself as a temporary regression. In addition, the effectiveness of a conversation attempt was a direct function of whether the subjects had anything to say to each other. For these reasons, the average data gives a clearer and more simplified view of the learning pattern.

On the other hand, the individual curves show the extreme differences in the kind of practice engaged in, as illustrated by the two examples of Graph 6.4b. The first pair of this figure, for example, stated explicitly that they were not engaging in phonemic practice since they felt that attempting conversations immediately would prove more fruitful. The other pair of this figure did more experimentation looking for some technique which would work, and did not have very much to say to each other once they were able to communicate. Spending a total of one hour practicing the names of cities, states, countries and books is indicative of this.

It must not be forgotten that the activities of the subjects are a measure of performance and not competence, even though competence is included in performance. Some of the pairs may have been able to



GRAPH. 6.4b PERCENTAGE OF TIME SPENT IN EACH STAGE FOR TWO PAIRS.

perform better, that is they were capable of better performance, but, for one reason or another, engaged in a simpler activity. The testing series is a better measure of competence, while the conversation sessions must be thought of more as the kind of learning practice each subject pair received. Another way of viewing the learning stages is to consider the aspects of the signal which are being attended to. Acoustic probing, and also phonemic practice, were activity focused on the spectral transformation medium. Concentrating on high-redundancy utterances composed of a set of limited phrases and words might be interpreted as learning to remember larger units of spectrally transformed speech. The synthetic conversations contained information transfer, and in this sense the subjects were listening for meaning, but they were using semantic redundancy of predicted content words to aid in perception. It might be said that during this stage subjects could understand transformed speech by concentrating on the utterance, but that the speed of natural speech exceeded their ability to process it. Hence, only some words were understood. As the process of perception became more automatic, true conversation characterized by a reduction of repetitions and semantic branching occurs.

The personality differences, mentioned in Section 6.2, appeared to show some differences in cognitive organization, in that one group of subjects treated the transformed speech as sounds and a second group treated it as the carrier of ideas. The former group spent considerable time using techniques for practicing, experimenting with the medium and speaking transformed speech, and in the extreme tried to generate phonetic dictionaries. The latter group tried, even at the beginning, to simply converse with each other. They attended to the thoughts and ideas of their partner using as much non-linguistic information as possible to enhance perception. Although their initial success was not significant, they persisted until they were able to have free conversations.

As stated in a guide for teaching foreign languages (Minnesota 1965), "no amount of theorizing, no amount of practice in fitting

individual parts together according to rule leads to proficiency in using language in actual situations...." It further points out that even though a form of practice may be mastered, this learning can not be transferred to the real situation. In rote learning it is the sequence being learned rather than the ability to automate the production and perception of new sounds. Dodson (1967 p. 84) also states that "the effort expended to imitate fluently and correctly the teacher's stimulus, diverts attention from establishing and consolidating a strong link between sentence meaning and the sounds he is speaking." Although he is referring to learning to associate meaning with production in the foreign language situation, this does not have to be the case. One can learn to understand a foreign language with a passive vocabularly just as one can learn to understand a dialect without being able to imitate it. Yet, learning must be considered an active process were what is being lerned is what is being practiced. Paradoxically, practice with transformed speech means listening to the ideas in a conversation and not speaking it. In fact, as will be shown in Section 6.5, those pairs who spent more of their time speaking transformed speech learned to understand it least.

6.46 LANGUAGE LEARNING SIMILARITIES

A set of analogies can be made between the four stages of learning transformed speech and those of learning a foreign language. In a typical foreign language classroom, the student is first taught the pronounciation of those sounds which are alien to his native language, and is concentrating on the perception and motor movements associated with the new sounds. As is the case in Stage I, very little attention is placed on word symbols, meaning, rhythm, idoms, etc. The instructor in the foreign language often gives the students some specific exercises in which they can practice the sounds. This might take the form of a limited vocabulary, a few sentences, a poem, or any kind of highly structured and limited linguistic text. The mastery of a select group of representative samples of the language does not allow the student to really communicate, but he does become accustomed to the language. Reciting a memorized poem is good practice even though it does not relate to the

ultimate objective--to communicate. This situation is typical of Stage II.

Once the student has mastered enough of the vocabulary, become familiar with the syntactic rules, and has the confidence to engage in conversation, he attempts to speak. If he were attempting this in the foreign country, he would find that communication was possible, but he would have to ask the native speaker to repeat sentences, or use other words if he did not understand. That is, he is understanding new words by categorizing and semantic branching. These activities are the characteristics of Stage III. Finally, after much practice, the foreign language is internalized, becoming automatic and releasing the cognitive attention for thinking about the ideas and not the words. True conversation occurs only when a person can "think" in the language. The same is true for transformed speech.

Similarly, analogies between the stages of learning and the process of acquisition of language can be drawn. Pre-language babbling is thought to be an activity oriented around acquiring motor control of the articulators and learning to correlate motor acitvity with the resulting sounds produced by the activity. This is essentially exploration of a speech medium in much the same sense that subjects experiment with the effects of the spectral transformation on their speech. When engaged in such activity, subject ignored the partner and concentrating on getting a feeling for sounds created by the system. The content of Stage I and babbling is non-linguistic.

According to Kaczmarek and others intonation is the first aspect of language which is learned by infants as young as 5 months (Weir 1966). He found that an infant's response to utterances which have the same intonation but different phonemic content was the same. Weir has shown that native speakers of a language can recognize an infant's babbling as being the same or different from the speakers' language. That is, the infant's babbling acquires the intonation characteristics of his language environment. In the language test,

described in Section 4.4, subjects recognized a passage consistently as being English or non-English*.

In an experiment described by Menyuk (1968), children were given meaningless sentences with scrambled word order, but with a natural stress pattern. Very young children, when asked to repeat them, reproduced utterances with the same stress pattern. Older children, however, tended to re-order the words to make a meaningful sentence and ignored the stress pattern. They were forcing meaning on to the content words. A very similar situation is found with the subjects' response to the sentences tests, described in Section 5.31. During the beginning of the experiment, subjects were only able to perceive some function words using, presumably, the prosodic features. As they became more experienced with the medium, the subjects perceived content words and created a sentence incorporating them, while ignoring the syntactic structure. Thus, they appear to make a transition from intonation-related structure to semantic content.

6.5 DIFFERENTIAL ABILITY OF SUBJECTS

Each pair of subjects created their own learning experience by choosing the content of the conversational sessions. Those subjects who were sensitive to the learning situation, designed efficient strategies for practicing transformed speech. In this sense, learning to understand transformed speech is a function of the content of the practice sessions and a function of the innate ability to master the new medium. It might be suggested that subjects would choose to practice material which they felt most competent to learn and they would therefore tend to improve on tasks relating to the content of their practice. Whether differences in learning strategies were a reflection of differences in verbal sensitivity, or vice versa, is impossible to say. Most assuredly, both factors are important. Even though cause and effect cannot be separated, some insight may be gained from studying the correlations between the amount of time spent in a given kind of learning

*Although some passages were incorrectly identified, the judgments were consistent from session to session.

practice and the performance on the testing series. Subjects were given rank socres based on the total amount of time spent in each stage for the first 12 conversational sessions. The Spearman correlation coefficient between these ranks and the ranks of the tests was then computed.

The total time spent in Stage I correlated very highly with the subjects' ability to identify vowels (R_s = .84, $p<.001$), but not consonants (R_s = .1). This result, when considered in terms of the inherent plasticity of the vowels (Section 4.24), implies that by practicing with sound patterns one can learn, or perhaps remember, the vowels, but this kind of learning does not enhance comprehension. The consonants, in contrast, are very difficult to learn, and practicing them does not improve performance. Moreover, practice with the acoustic medium did not enhance perception of words* or sentences. The time spent learning to speak or imitate spectrally transformed speech correlated negatively with the performance score for comprehending sentences and identifying the names in the geographic-names test** (R_s = -.64, $p<.05$ for sentences; R_s = -.58, $p<.05$ for geographic names). That this correlation is significant and negative demonstrates the basic and inherent differences between perception and production and, therefore, tends to weaken the anlaysis-by-synthesis theoreis of Stevens (1960) and, Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967). In these theories, it is assumed that the cognitive mechanisms involved in speech production are also used in speech perception. If the analysis-by-synthesis, or the motor theory of perception, model were valid, then one would expect that learning to speak spectrally transformed speech should enhance perception and vice versa. That is, the cognitive mechanisms used to

*Some of the word tests correlated with acoustic practice; however, the correlations were not consistent and showed no general pattern. For example, phonemic practice correlated with two syllable words (R_s = .80) but did not correlate with one syllable words (R_s = .16). The difficulty is caused by the very low raw scores for the word tests which makes the rank scores unreliable.

**As discussed in Section 5.4, it is felt that the same cues are used for comprehending sentences and identifying the geographic names.

generate speech are thought to be used to perceive speech and, consequently, learning to speak transformed speech must allow for perception of transformed speech since the same mechanisms are used for both.

Similarly, the time spent in Stage II, high-redundancy utterances, correlated negatively (but not significantly) with sentence comprehension, consonant identification, and word perception. In summary, practicing isolated sounds or experimenting with the medium never correlates positively with comprehension; practice with vowels can improve vowel perception but does not result in improved comprehension.

Stage IV, free conversation, correlated with the subjects ability to comprehend sentences and identify geographic names with R_s = .68 and R_s = .61, respectively ($p<.05$). This is to have been expected since conversation is simply made up of sentences. A subject who can converse must be able to understand sentences, Parenthetically, Stage IV also showed the only positive, significant correlation with the ability to identify consonants (R_s = .58, $p<.05$). This somewhat interesting result must be interpreted with caution. On one hand, it might be said that free discourse eventually leads to an improvement in consonant perception or that experience with a varied context teaches the subjects how to extract the rather subtle consonant cues from the neighboring vowels. Those subjects who spent more time successfully conversing were exposed to more transformed speech. But on the other hand, the ability to converse might be the result of the subjects' innate sensitivity to the consonants. Since consonants carry more linguistic information than vowels, those subjects who could better perceive consonants should find conversing much easier than those who had trouble with the consonants. To answer this question of cause and effect, a special statistical test was created.

The statistical measure D is equal to the sum of the magnitude differences in ranks for each of the subject pairs on a particular test. If the subject pairs are randomly matched for a given ability, then one would expect that a subject with a low rank is equally likely to be matched with one who has either a high or low rank. But, if the

ability is a reflection of the kind of practice which the two subjects of each pair engaged in, then one would expect the ranks of the subject pairs to be more similar than chance. In essence then, D is a measure of the similarity in performance, on the average, as a result of the fact that the conversational practice was unique to each pair. Using this measure the following question can be answered: Was learning the natural result of exposure to the transformed speech medium or was it the result of the kind of practice which the subjects were engaged in? Moreover, if the subjects reach a level of performance determined by their native ability, then one would expect that the D would be insignificant since the subjects were not matched for any ability. See Appendix C for further details on the statistic D.

When this statistic is used on the consonant identification ranks it is found that D decreases from 34 (maximum is 36) on the first session to 11 (minimum is 6) on the last session. Thus, on the initial base-line test it is found that the subject pairs are actually mismatched ($p<.07$) for their ability to correctly identify consonants, but by the end of the experiment they are almost completely matched ($p<.005$). This highly unambiguous result unequivocally indicates that the kind of practice is the determining factor in the ability to learn consonants, and that conversing rather than phonemic practice is the best method for learning how to use them in perceiving natural conversation. The importance of this conclusion cannot be overstated for it emphasizes the point that comprehension learning arises from integration of automatic recognition. It does not arise from learning the segments; segment learning does not transfer to total perception required for speech.

Similarly, the D for identifying vowels correctly decreased from 23 on the first session to 12 on the last session ($p<.005$). Most of the decrease in the D took place immediately after the first session, which corresponds to the learning curve of Graph 4.2b. The correlation between vowel performance and the time spent in Stage I is also the result of the kind of practice engaged in.

In contrast to the consonant and vowel identification, the D for the ABX consonant discrimination remained approximately constant, but with an average value of 31 ($p<.15$). This means that the subjects were mismatched in their discrimination ability and that the learning shown in Graph 4.3a is not a function of the kind of practice, but rather it is a characteristic of the subjects.

CHAPTER VII

CONCLUSIONS

7.1 SUMMARY

The electronic transformation system used in this experiment rotated the bandlimited spectrum of the incoming speech wave around the center rotation frequency of 1600 Hz. Thus, the spectral envelop, as well as the individual harmonic components, of the acoustic wave were inverted so that the high frequency energy became low frequency energy and vice versa. In terms of the excitation to the auditory nerve, the components, which previous to the transformation, stimulated the cells near the stapes, now stimulated the cells maximally distant from the stapes. Essentially, then, the neural firing pattern for the lowest level auditory cells is also reversed. In the frequency region of interest, the psychophysical characteristics of the auditory system are approximately constant, although there appear to be different modes of response above and below 1000 Hz. Nonetheless, one would predict that the transformation does not result in an excessive loss in hearing sensitivity.

The subjective sensation of pitch, which is related to the frequency difference between the harmonic components and the predominant envelop modulation, was unaffected by the transformation. Also, stress, loudness, duration, juncture, noise, silence, etc. were spectrally independent and therefore perceived normally. Only those speech features that are a function of the formants and formant transitions were modified so that they could not be initially perceived.

The ABX discrimination tests showed that two distinct phonemes generally remained distinguishable after transformation, although they may not have been identified correctly. Only discriminations based on the place of articulation feature for the consonants were difficult to make. In particular, the error rate for the place feature was about 25% for all consonants and 44% for the unvoiced plosives by the end of the

experiment.

From the results of the phoneme identification tests, it was found that several of the distinctive features have an acoustic manifestation which was clearly perceived, and that these features were spectrally independent. Very seldom were there any confusion errors between phonemes that differed in the tense-lax, voicing-nonvoicing, and mode of articulation features. Initially the errors in these features were small, and by the end of the experiment they were almost nonexistent. The spectrally dependent features showed a very different kind of behavior. In the base-line test of session 1, phonemes characterized by low-frequency energy were perceived as those characterized by high frequency energy. This was particularly true of the tense vowels, but only somewhat true of the consonants and lax vowels. After a minimal amount of exposure to the transformed medium, the situation reversed itself. Now, the tense vowels were perceived correctly and the lax vowels were more easily identified. Subjects never, however, gained the ability to identify the place of articulation feature for the consonants even though some improvement occurred.

The asymptotic score for the identification of vowels was about 30% (chance equals 9%), which occurred around session 3, and for the consonants it was about 35% (chance equals 7%), which occurred around session 8. Beyond this point no further improvement resulted from increased conversational practice.

It was discovered from the word comprehension tests that subjects had a great difficulty in correctly perceiving a word without any extra cue. The scores for the one- and two-syllable word tests remained small throughout the experiment because many of the words used were less common than others within the same confusion class. However, when the subjects were told that the words were the names of cities, states and countries, their score was very high, 27% on the first session and 70% on the last session. Similarly, when they were told the category from which the words were taken their scores were much higher

than when they were not told the categories, 58% and 8% respectively. Words embedded in a sentence were also much easier to perceive than isolated words, but a carrier sentence which provided an acoustic context without providing any semantic redundancy did not enhance perception.

The sentence comprehension tests demonstrated that knowing the theme of a group of sentences increased intelligibility by approximately a factor of four over what it would have been without the extra cues. Nonetheless, by the end of the experiment the average score for random sentences was 35%. It is of great consequence to note that the major improvement in sentence comprehension and successful dialogues between subject pairs occurred after improvement of phoneme identification had ceased. That is, even though the subjects showed no improvement on the isolated language tasks, they showed very pronounced increase in their ability to comprehend normal utterances.

The error pattern on the sentence tests of session one revealed that subjects were able to perceive isolated function words embedded in continuous speech without understanding any of the content words. The ability to identify such words as "and", "the", and "is" must be attributed to the relationship between prosodic features and the syntactic structure. Once they began to understand some content words they no longer relied on their ability to follow the stress pattern and would manufacture a sentence around the content words.

Owing to the highly individualistic nature of the conversation practice and the natural differences in personality, the scores on the various tests showed considerable variation. This variation was exploited in determining correlations between the different abilities. It was found that the ability to identify consonants, identify vowels, discriminate phonemes, and discriminate languages were all uncorrelated with each other. Although the ability to perceive consonants did correlate with the ability to comprehend words, word performance did not correlate with sentence comprehension. However, the rank scores for the identification of the names of cities, states and countries

correlated very highly with sentence comprehension, thus showing that the mechanisms involved are probably similar. In both cases, the listener used cues relating to syllable length, semantic redundancy, stress, etc. The lack of most correlations must be interpreted as meaning that the performance on the individual tasks is not necessarily a reflection of the kind of cognitive processing which occurs under normal language perception.

The conversational learning sessions were categorized by labeling the subjects' activity as being one of four stages: acoustic probing, high redundancy utterances, semantic branching, and free conversation. In the initial stages, subjects explored the medium using phonemes, non-sense sounds, and other exploratory utterances. After this kind of acoustic probing, they began to communicate using a very limited set of words and ideas, as illustrated by the use of spelling. Semantic branching is characterized by a strategy of communication revolving around verbal associations and alternative ways of saying the same thought. Free conversation is simply communicating with a minimum of repetitions and semantic branching. In the discussions of the various stages analogies were drawn between learning to understand spectrally transformed speech and foreign language learning. Also, analogies with the stages of children's language learning were made.

The amount of time that each subject spent in a particular activity often correlated with his performance on the tests. The total time spent in acoustic probing correlated with vowel identification but not with consonant perception. Phonetic practice sometimes correlated with the ability to perceive isolated words, but comprehension of sentences correlated negatively with the amount of time spent in trying to speak transformed speech. The only ability which correlated with sentence comprehension was the time spent in free conversation. Interestingly, the more time spent in conversation, the better the subjects' ability to identify consonants. That is, practicing with phonemes is not good for learning consonants as is simply conversing.

7.2 THEORETICAL IMPLICATIONS

7.21 PLASTICITY AND THE VISUAL ANALOG

The initial motivation for this spectrally rotated speech experiment came from observing the apparently analogous visual experiments using inverting, rotating, and shifting prisms. Both the rotating prisms and the spectrally rotating electronics operate on the stimuli in order to transform the neural excitation pattern. Viewed in this way, the analogy is accurate since there is a point-to-point correspondence between frequency and position along the cochlea, just as there is between vertical space and the position along the retina. It has been thought that adaptation to the visual transformation demonstrated a kind of plasticity in the perceptual mode, in that the cognitive system compensates for the distortion so as to restore the correct perception. An explanation of this kind requires extreme caution when one defines plasticity of perception. The interpretation of the visual transformation experiments has been debated for the last half century following the classic experiment of Stratton (1897). Recently, Harris (1965) has challenged the very notion of plasticity, compensation, and inverse transformation by suggesting that the results of the prism experiments mean that the subjects have adapted to the change in the relationship between their vision and their body sense. That is, their vision is not changing but the use of vision to form a sense orientation and a relationship to the environment is being modified. The wearer of inverting prisms does not "see" the same world he would "see" without the prisms, but he learns to orient his visual image of his body so that it agrees with the world.

If this in fact is the case, then, one is faced with a very difficult question with regard to the transformed speech experiment. What are the subjects learning when they have learned to communicate through the medium? The auditory system, unlike the visual system which is used for orientation information, functions primarily for the transfer of acoustically encoded verbal information. It might be argued that auditory feedback is used to maintain correct pronounciation,

but it was found that the subjects, articulation remained unaffected when they were listening to their own voice transformed. Although delayed auditory feedback can seriously impair speech (Fairbanks 1966, p. 11; Lee 1950), other forms of distortion which leave the temporal sequence intact do not affect production of speech. If anything, then, feedback is used for timing.

Before asking the question, "what have the subjects learned?", one must ask about what has been changed or removed by the spectral transformation. Only if one views the input to the perceptual system as a time varying frequency spectrum pattern exciting a pattern recognition system, as Stevens (1960) has in his model of speech recognition, must one conclude that the transformed speech is a completely new pattern. Other forms of distortions, such as clipping, do not remove the spectral pattern but only obscure it by adding extraneous harmonics. This distinction is apparent rather than real. Our present knowledge of pattern recognition and our heavy emphasis on speech spectrographs in speech analysis has created a false impression. Perception, being an active process, does not depend on the stimulus to the extent that any passive pattern recognition system would. If perception were based on the detection of a single set of features, then one could calculate how much of the pattern remains after the speech has been spectrally inverted. However, one may take the view that there are multiple sets of cues in the acoustic signal and that removing one set does not necessarily destroy the speech; in order to learn to understand transformed speech, the listener must create new cognitive strategies which can use other sets of cues. Thus, the subjects in this experiment are never learning to re-invert the spectrum, nor are they learning to use the cues which they used with natural speech.

Learning to use the more subtle frequency cues in spectrally transformed speech is probably very similar to learning to converse in a foreign language without an accent. This usually requires extensive practice, a language aptitude, and exposure to the language at a young age.

Two good examples illustrate the inadequacies of spectral distortion for destroying intelligibility. Denes (1964) carried out an experiment in which the speech spectrum from 180 to 4500 Hz was compressed into a range from 50 to 1600 Hz using a modified vocoder excited with 1/3 of the original pitch. Over the course of 16 sessions the word comprehension increased from about 40% to 70%. With additional contextual information, sentences would be almost completely intelligible.

A privacy system used during the early 1940's divided the speech spectrum from 200 to 3200 Hz into 5 equal bands each of which could be shifted to a new frequency and rotated (Schott 1943). Of the 3,840 possible combinations of the rotated and shifted bands, only 11 were found suitable for privacy. All combinations with one band in the correct place or with two neighboring bands in the correct order were ruled out. Tests given to two trained subjects using one of the 11 special codes showed an average articulation score with sentences of 41% and a maximum of 81%. When the code combinations were switched at a rate of .06 seconds per code the intelligibility was about the same as for a fixed code. Since the codes were changing, comprehension can not, under any condition, be attributed to learning the transformation. The subjects had to be using cues that remained unaffected by spectral distortion. The results also showed that there are enough of these cues available to a listener who knows "how to perceive" them to enable him to understand this kind of speech.

The spectrally independent features are the same as those mentioned in Section 4.1, namely the temporal pattern of loudness, pitch, and duration, as well as the mode of articulation, voicing and tense-lax feature. However, in the same set of experiments, Schott reports that when a time scrambler interchanged ten segments of about 37 milliseconds the average articulation score was 63% with a maximum of 87%. The temporal distortion alone, then is also not sufficient to destroy intelligibility since the listener can use the spectral cues for perception. A combination of the time and frequency scrambler, however, was able to reduce the intelligibility to 10%. Thus, only by destroying

both the time and frequency cues could the speech information be removed from the acoustic signal.

Amplitude distortion, such as peak clipping, is probably the least effective form of information destroying processes. Licklider and Pollack (1948) found that differentiating followed by infinite peak clipping was almost completely intelligible. Similarly, interrupted speech is also intelligible unless the switching rate is equal to some function of the syllable rate (Miller and Licklider 1950; Huggins 1964).

These examples should be sufficient proof that cognition is not simply the result of recognizing a pattern, but that it is an active process which can use whatever cues are available as a basis for perception.

7.22 COGNITIVE PROCESSING

Speech is an encoding of phonemic units onto an acoustic carrier; and therefore, perception must be viewed as the decoding process for the extraction of the phonemically represented message. The way in which the decoding takes place is the essence of perception. A rather simple argument demonstrates that the decoding cannot be based on phonetic or phonemic units. Since speech can be understood at rates as high as 400 words per minute (Orr, Friedman and Williams 1965), with a corresponding 30 phonemes per second, the listener would have to be making a phonetic decision every 30 milliseconds. Not only can he not identify a single phoneme sound of this length, but a continuous stream of independent sound units would merge into an unanalyzable buzz (Miller and Taylor 1948). The unit of perception must, then, be larger than the phonemic units.

In complete opposition to the sound unit idea, Miller (1962) has suggested that the decision units are phrases with as many as three words. He reasons that a cognitive unit involves a decision and that studies of reaction time would give evidence to 200 milliseconds as being the length of time needed for such a decision. Furthermore,

each decision would be choosing one alternative among a vast variety of possible phrases, but the reaction time is independent of the number of possible alternatives if they are highly familiar. The experiments of Fodor and Bever (1965) and Garrett, Bever and Fodor (1966) also suggest a phrase-related unit of perception. They have shown that, when a click is presented concurrently with a sentence, perception of it tends to be shifted to the phrase structure boundary. The implication is that the listener is processing the decision unit and that only afterwards does the click become apparent. Huey (1968, p. 72) reports, in the case of visual tachistoscope experiments, that "it was found that when sentences or phrases were exposed, they were either grasped as wholes or else scarcely any of the words or letters were read."

The problem of speech perception thus reduces to that of deciding on the correct phrase element in the time available. With transformed speech, the listener must make the same kind of decisions that he makes with normal speech, except that the cues which are available are somewhat different. The question "What have the subjects learned when they understand transformed speech?" becomes two distinctly different questions "What are the cues and features he learns to use?" and also "What algorithms does he use for organizing the cues?". By viewing the problem in this way, one realizes that the results of the testing series, which showed no additional learning beyond session eight, do not conflict with the observation that sentence comprehension improvement occurred mostly after session eight.

The phonemic test results, described in Chapters IV and V, showed that the subjects were never able to learn to distinguish the place of articulation feature for consonants and lax vowels, but aside from this, they were able to attain almost perfect scores on the other speech tasks. Since they could also perceive stress and other prosodic features, one would expect them to be able to understand normal speech. As Miller and Nicely (1955) point out, elliptic speech is quite intelligible. Yet, true comprehension occurred towards the end of the experiment and only a few subjects reached that stage. Our explanation, then, is not complete unless we include the fact that it is during the

second half of the experiment that the subjects are trying to learn to use the new cues rapidly enough to follow the normal conversational speech rate.

Based on this theoretical point of view, it is felt that the rate of cognitive processing is directly related to the "break-through" phenomenon mentioned in Section 6.43. Preceding break-through, subjects were learning to use new cues required for the perception of phonemic units, acquiring a familiarity with the effect of the medium on continuous speech, developing a facility to use the prosodic features for segmentation and integration of the utterance, but the fact that these abilities were not sufficiently over-learned prevented them from comprehending continuous speech.

It has been demonstrated in this and other experiments that contextual information is a very significant factor in the perception of speech. Notice, however, that contextual information is gained by understanding the previous utterance or part of the same utterance. Thus, if the listener is processing the transformed speech at a rate slower than the natural speech rate, he will fall behind and lose the content information which would have made perception of the next speech unit easier. Hence, frequent repetitions are necessary in order to communicate a utterance longer than one phrase unit. After his processing of the transformed speech features has become sufficiently automatic, by developing strong cognitive links between verbal units and the acoustic cues, he can follow the speech without falling behind. At this point, contextual information is continuously available, so that perception and comprehension suddenly increase. The learning interval in which the processing rate becomes equal to the speech rate is thought to be directly realted to the sudden improvement characteristic of "break-through". The listener is able to "lock" on to the content.

This hypothesis also explains why subjects found it very difficult to change the topic of conversation. When all utterances related to a common theme they had ample contextual information; switching topics, however, required that they establish the new topic and a new

set of contextual constraints. Cherry (1957, p. 276; also Chapter 5.32) observed that subjects, when told the theme of a discussion, were able to understand speech in the presence of excessive masking noise, but were not able to understand it without the content cue.

If the above discussion is correct, then the effects of spectral transformation on comprehension cannot be viewed in the same way as the effects of other distortion, filtering, and additive noise experiments. In those kinds of experiments, very little learning occurs, as shown, for example, by word comprehension curves of subjects listening to clipped speech (Licklider and Pollack 1948). Unfortunately, very little data appropriate to this question are available. In an investigation of articulation testing methods for different kinds of systems, Williams, Hecker, Stevens, and Woods (1966) demonstrated that when the effects of familiarity with the test stimuli are removed very little learning takes place. They conclude that a closed-set intelligibility measure is an accurate reflection of the system and is independent of other factors, which presumably includes learning.

The break-through phenomenon is, perhaps, then, more closely analogous to the circumstances in learning a foreign language at the point in time when the listener knows the vocabulary and the syntactic rules of the language, but is unable to comprehend a complete utterance. By concentrating on a single word in the utterance, he gains enough time to perceive it, but he has fallen behind the sequence of words following it. It is perhaps for this reason that language teachers often give repetition exercises to increase the proficiency, which when translated into psychological terms, means learning to perceive more rapidly.

7.23 EXTENSION OF ANALYSIS-BY-SYNTHESIS MODEL

The speech perception model most currently in vogue, even though there are some criticisms of it (Lane 1965), is the analysis-by-synthesis model (Halle and Stevens 1962; Stevens and Halle 1967) and a modified version thereof based on a motor theory of perception (Liberman,

Cooper, Shankweiler, and Studdert-Kennedy 1967; Liberman, Cooper, Harris, MacNeilage, and Studdert-Kennedy 1967). In these models, the incoming acoustic speech wave undergoes a preliminary analysis and is then compared to some transformation of the listener's hypothesis about the utterance. The differences between the listener's hypothesis and the speech stimulus are used to correct and improve the hypothesis until a successful match is found. Such a model, although appropriate in many ways, cannot explain how a listener can understand a severe accent or dialect, how he can listen to one voice in the presence of many others, or how he can learn to understand spectrally transformed speech. If the comparison between the hypothesis and the incoming signal is based on a similarity of a set of features which, for example, might include phonetic cues based on formants, then no match between transformed speech and the internal hypothesis could be made. The definition of similarity is the crux of the matter. To illustrate this, consider that the spectrally transformed vowel /i/ is more similar to the untransformed /u/, using most any physical basis of similarity, yet it is perceived correctly as /i/ after half an hour of practice.

The model shown in Fig. 7.2a is an extension of the basic model of Halle and Stevens, but it includes a number of extra variables to allow the definition of similarity to be made by the listener. The incoming acoustic wave undergoes a preliminary analysis which reduces it to a set of phonetic features. The feature set may, although not necessarily, be related to our present notion of distinctive features and is sufficiently compact to allow a reasonable large segment, say 300 milliseconds, to be held in temporary storage. Based on the listener's hypothesis of what is being said, a comparison between the actual and expected features is made. The errors from the feature comparisons are then evaluated by the box labeled "evaluator and decision maker" which uses one of many possible algorithms for interpreting similarity.

Thus, for example, if the listener is interested in phonemes rather than semantic content, as a phonetician would be, he inputs a

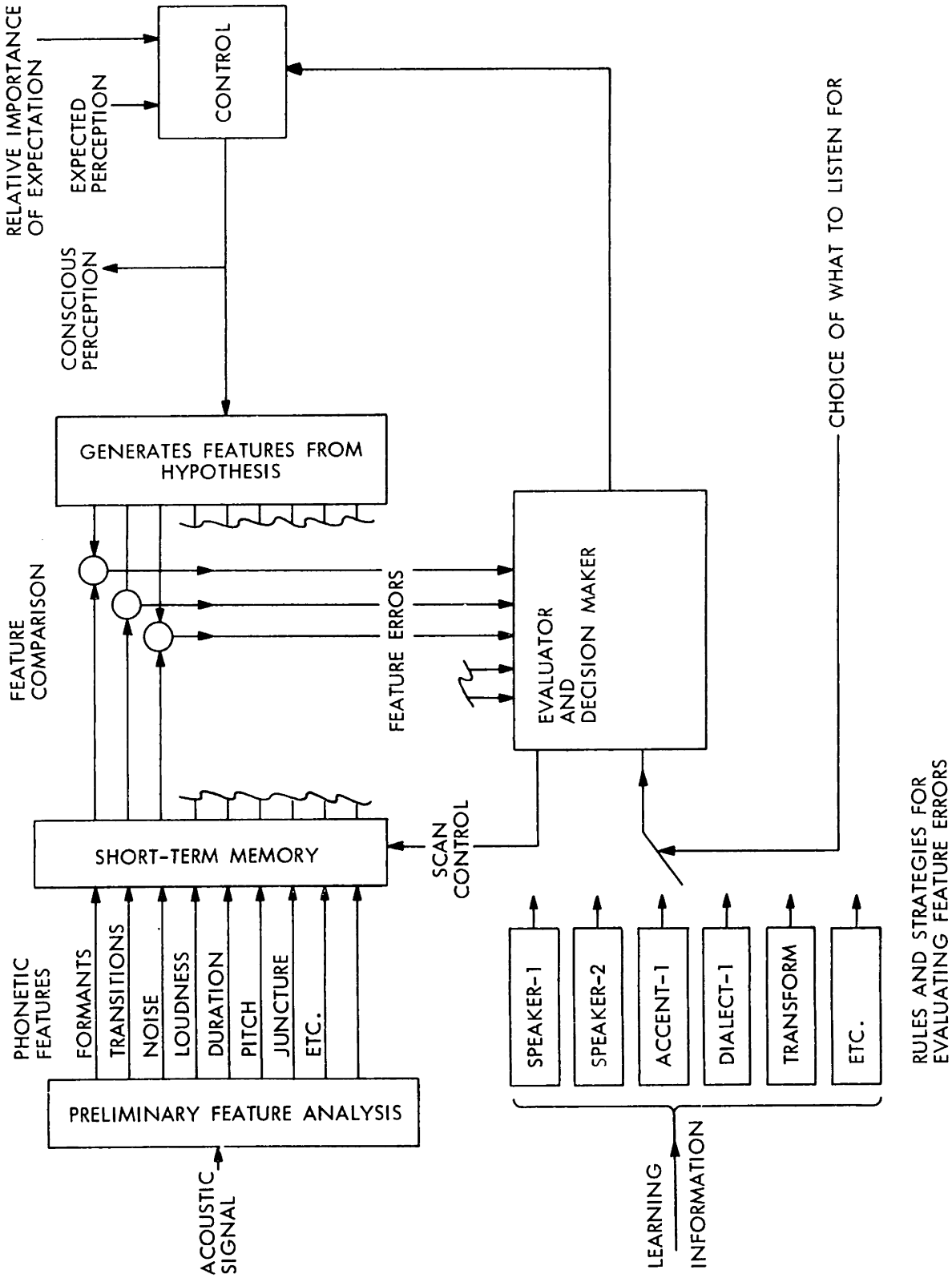


FIG. 7.2a SPEECH PERCEPTION MODEL.

set of strategies into the "decision maker" which force a match to be made solely on the basis of the acoustic properties of the sound units, and ignores contextual and content information. In the normal situation, the listener uses a set of rules which take into account the idiosyncratic aspects of the speaker's speech, as for example, an accentuated /y/ glides on front vowels or incorrect place of articulation of /w/ for lisping. These cues, in addition to differences in pitch and formant range, allow a listener to concentrate on only one speaker in the presence of many other voices. Also, in a situation where the listener is not paying close attention, he is instructing the "control" box to place more importance on the expected utterance ignoring the decision-maker evaluation.

The learning process is, then, the acquisition of an appropriate set of strategies and rules for the proper evaluation of the differences between the internal hypothesis and the incoming signal. For transformed speech, the rules might include such factors as: ignore formant transition for some consonants, emphasize the importance of the prosodic features, expect large errors for tense vowels, etc. Moreover, in the initial learning stage, the control unit, which actually generates the hypothesis, might be instructed to place more importance on the expectation based on the previous context and less importance on the lack of a successful match. Once the listener has learned to use the rules he can increase his dependence on the decision maker. Notice, that in the model the decision maker controls the scanning of new data into temporary storage. Thus, if the inexperienced listener is spending too much time evaluating a given segment of the utterance and not scanning the data at the speech rate, he will fall behind. In effect, then, he has the choice of either listening carefully to some of the segments or allowing the system to proceed at a normal rate making inconclusive judgments about what was being said. This is a free choice, as is the reliance on previous context. Once the entire process has become automatic, assuming the subject has found a good set of rules to use with the transformed speech, he can relax and allow the perceptual system to comprehend the speech as if it were undistorted.

7.3 PRACTICAL IMPLICATIONS

7.31 SPEECH SYNTHESIZERS

Perhaps the most relevant conclusion of this experiment is that the temporal prosodic features can be more important than phonetic cues. The design of speech production algorithms seems to emphasize the importance of finding a set of parameters for specifying the phonetic components of the speech wave, rather than trying to incorporate a set of global parameters which relates syntactic and semantic content to the prosodic features. It would appear that effort spent at implementing a stress pattern, even crudely, would be more beneficial than improving the characteristics of the phoneme sequence. In this experiment, no subject was ever able to identify a /p/, /t/, or /k/, for example, yet many subjects learned to converse with each other. A major difficulty in understanding some speech synthesizers results because the listener segments the continuous acoustic wave in the wrong places. This problem is avoided with the appropriate juncture and stress.

The lack of a correlation between phoneme performance and sentence comprehension must be interpreted as meaning that speech synthesizers should be evaluated with complex stimuli. It is conceivable that a synthesizer which could produce easily recognized phonemes in isolation would not be able to produce continuous speech. Alternatively, recognizing phonemes in a sentence context is not completely dependent on the phonetic characteristics used to generate the phonemes.

7.32 SPEECH RECOGNITION

The question of recognizing speech is far more subtle than producing it since one is faced with the infinite variety of idiosyncratic speech patterns. Moreover, a human listener recognizes speech by using many levels of cognitive processing, starting from the highest level of expected meaning and working down to phoneme level when necessary. The fact that listeners can understand spectrally transformed speech means that the acoustic signal, per se, is not directly related to the perceptual response. No schematized speech recognition system, other than

a human being, could possibly learn to equate the acoustic signal of the transformed and untransformed speech. The lack of a direct dependence on the acoustic signal is an ill omen for speech recognition modeled on human perception. Unless the system were to include a semantic dictionary, a syntactic grammar analysis, and the human experience of living, it could not parallel the behavior of the human being. Only under the most ideal condition can the signal and the phoneme sequence be related.

Moreover, if one were to try to model the more complex cognitive process of speech perception, and if one could succeed, one would probably find that such a system would not work in real-time and that it would not work very well at all. Perhaps, then, the best initial approach would be to disregard the relevance of cognition in trying to build such recognition systems and, rather, orient them around the available devices. A digital computer can perform simple operations in an extremely short time, an addition, for example, in less than 1 μsec; but it could not look a word up in a dictionary of 10,000 words to find the right entry in 100 milliseconds. The listener, however, is essentially extracting meaning and function from a dictionary at a rate approaching one word per 100 milliseconds. The problem thus reduces to clever ways of organizing the computer algorithms to allow for complex processing in a short amount of time. The answer does not lie in imitating a human cognitive system since the brain is, according to our present understanding, a performing parallel rather than serial operations. The use of a computer restricts the alternatives far more than our limited knowledge of speech perception does.

APPENDIX A

CONTENT OF TESTS

LANGUAGE DISCRIMINATION TEST

SESSIONS 1, 20

Arabic, Chinese, Japanese, English (1), Russian, Hindi, English(2), Rumanian, French, German, English(3), Finnish, Hebrew.

SESSIONS 4, 13

Chinese, English(3), Russian, Finnish, English(1), Chinese*, French**, English(2), Rumanian, Arabic, German, Hindi, French.

SESSIONS 7, 16

Hindi, Rumanian, English(2), French, Arabic, Chinese, German, English(3), Japanese, Finnish, Hebrew, Russian, English(1).

SESSIONS 10, 19

Arabic, Finnish, Hebrew, English(1), French, Rumanian, Chinese, Hindi, English(2), German, Russian, English(3), Japanese.

*Should have been Hebrew

**Should have been Japanese

VOWEL AND CONSONANT DISCRIMINATION TESTS

SESSIONS 1, 20

met, met, met
week, work, week
bed, bed, bird
use, us, us
few, fur, few

hurt, hurt, hut
hide, hide, hid
so, saw, saw
goat, goat, got
tow, to, tow

broke, broke, brook
short, shot, shot
draw, draw, drew
talk, talk, took
hop, hoop, hop

God, good, good
full, fool, full
ox, oaks, oaks
pull, pull, pole
they, they, there

rain, ran, rain
make, mark, mark
grass, Grace, Grace
take, talk, take
fair, fair, far

am, am, arm
cat, caught, caught
aunt, ant, aunt

SESSIONS 1, 4, 13, 20

pie, pie, by
curl, girl, girl
sit, fit, sit
tea, key, tea
mice, mice, nice

rise, rise, wise
fan, van, van
vine, thine, vine
she, she, fee
think, sink, think

pan, pan, tan
ring, rim, ring
yoke, woke, woke
thin, thing, thing
vat, bat, bat

thigh, thy, thy
see, zee, see
tail, tail, sail
fair, pear, pear
do, zoo, do

tore, door, tore
wet, let, let
call, Pall, Pall
rule, rule, yule
bed, bed, dead

bet, get, bet
light, right, right
then, then, Zen

SESSION 2

tow, to, tow
broke, broke, book
full, fool, full
pull, pull, pole
ate, oat, oat
show, shoe, shoe
men, man, men
will, well, will

tow, tow, to
brook, broke, broke
full, full, fool
pole, pull, pull
oat, ate, oat
shoe, show, shoe
men, man, man
will, will, well

to, tow, tow
broke, brook, broke
fool, full, full
pull, pole, pull
oat, oat, ate
shoe, shoe, show
man, men, man
well, will, will

to, tow, to
brook, brook, broke
fool, full, fool
pole, pole, pull
oat, ate, ate
shoe, show, show

SESSIONS 1, 20

air, air, ear
may, my, may

ate, oat, oat
day, do, day
may, mew, mew
meet, meet, might
see, so, so

he, who, he
knee, knee, new
kite, kite, coat
tie, to, to
eyes, use, eyes

show, shoe, shoe
man, man, men
an, in, in
will, well, will
back, book, book

ran, run, ran
ax, ax, ox
get, got, get
fell, fell, full
beg, bug, bug

SESSIONS 1, 4, 13, 20

fell, hell, hell
yes, less, yes

gate, gate, date
said, shed, said
so, so, though
thin, shin, thin
then, hen, hen

thick, thick, hick
three, tree, three
foe, though, foe
think, fink, think
hick, hick, vick

they'll sale, sale
nine, nine, line
mice, vice, vice
vile, vile, gile
bog, bong, bong

ball, fall, ball
more, fore, fore
best, mest, best
zoo, zoo, Vu
pet, vet, vet

SESSION 2

men, men, man
well, will, well

to, to, tow
broken, brook, brook
fool, fool, full
pull, pole, pole
ate, oat, ate
show, shoe, show
man, men, men
well, well, will

tow, to, to
brook, broke, brook
full, fool, fool
pole, pull, pole
ate, ate, oat
show, show, shoe
men, man, men
will, well, well

tow, to, tow
broke, broke, brook

SESSIONS 7, 16

vet, pet, vet
Vu, zoo, zoo
best, best, mest
fore, more, fore
ball, ball, fall

SESSIONS 10, 19

get, bet, bet
right, right, light
then, Zen, then
hell, hell, fell
less, yes, yes

SESSIONS 7, 16

bong, bog, bong
gile, vile, vile
vice, mice, vice
line, nine, nine
sale, they'll, sale

vick, hick, hick
think, think, fink
foe, foe, though
three, three, tree
hick, thick, thick

hen, then, hen
thin, thin, shin
though, so, so
said, said, shed
date, gate, gate

yes, yes, less
hell, fell, hell
Zen, then, then
right, light, right
bet, bet, get

dead, bed, bed
yule, rule, rule
Pall, call, Pall
let, wet, let
door, tore, tore

do, do, zoo
pear, fair, pear
sail, tail, tail
see, zee, see

SESSIONS 10, 19

gate, date, gate
shed, said, said
so, though, so
shin, thin, thin
hen, hen, then

thick, hick, thick
tree, three, three
though, foe, foe
fink, think, think
hick, vice, hick

sale, sale, they'll
nine, line, nine
vice, vice, mice
vile, gile, vile
bong, bong, bog

fall, ball, ball
fore, fore, more
mest, bets, best
zoo, Vu, zoo
vet, vet, pet

pie, by, pie
girl, girl, curl
fit, sit, sit
key, tea, tea
mice, nice, mice

rise, wise, rise
van, van, fan
thine, vine, vine
she, fee, she

SESSIONS 7, 16	SESSIONS 10, 19
thy, thigh, thy	sink, think, think
bat, bat, vat	pan, tan, pan
thing, thin, thing	rim, ring, ring
woke, yoke, woke	woke, woke, yoke
ring, ring, rim	thing, thing, thin
tan, pan, pan	vat, bat, bat
think, think, sink	thy, thy, thigh
fee, she, she	see, see, zee
vine, vine, thine	tail, sail, tail
van, fan, van	pear, pear, fair
wise, rise, rise	zoo, do, do
nice, mice, mice	tore, door, tore
tea, tea, key	let, let, wet
sit, sit, fit	Paul, Paul, call
girl, curl, girl	rule, yule, rule
by, pie, pie	bed, dead, bed

VOWEL AND CONSONANT MATRIX TESTS

SESSIONS 1, 20

ka, fa, ta, za, ba, ga, sa, ma, sha, θa, ja, pa, tha, na, da, va, sa, ma, za, ja, ta, pa, da, sha, na, fa, ja, tha, va, θa, ka, ba.

SESSIONS 2, 11

pa, va, na, ba, za, ta, sa, fa, ja, ba, sha, ma, na, ga, va, da, θa, ka, sa, tha, ta, da, za, ga, pa, θa, ka, tha, fa, ma, sha, ja.

SESSIONS 5, 15

sa, ma, za, ja, ta, pa, da, sha, na, fa, ga, tha, va, θa, ka, ba, ka, fa, ta, za, ba, ga, sa, ma, sha, θa, ja, pa, tha, na, da, va.

SESSIONS 8, 17

θa, ka, sa, tha, ta, da, za, ga, pa, θa, ka, tha, fa, ma, sha, ja, pa, va, na, ba, za, ta, sa, fa, ja, ba, sha, ma, na, ga, va, da.

SESSIONS 1, 20

boot, bet, bought, Bert, but, bat, bit, bate, boat, beet, bite, boat, boot, bite, bat, bet, bit, Bert, but, beet, bate, bought.

SESSIONS 3, 12

bate, bought, bite, but, boat, beet, bat, bit, boot, Bert, bet, bought, bate, bite, but, bit, bat, boat, beet, boot, Bert, bet.

SESSIONS 6, 15

bet, bate, bite, bit, bought, Bert, boat, boot, bat, but, beet, bite, bet, bate, bit, bought, Bert, bat, boot, but, beet, boat.

SESSIONS 9, 18

Bert, but, bite, bit, boot, bate, bet, bat, beet, boat, bought, Bert, bit, but, bat, bite, boot, bate, bet, bought, beet, boat.

UNVOICED PLOSIVE TEST

SESSIONS 1, 20

pole, coal, toll	pole, coal, toll
peas, keys, tease	peas, keys, tease
pan, can, tan	pan, can, tan
Paul, call, tall	Paul, call, tall
pen, Ken, ten	pen, Ken, ten
pool, cool, tool	pool, cool, tool
pill, kill, till	pill, kill, till
pave, cave, tave	pave, cave, tave
puff, cuff, tough	puff, cuff, tough
par, car, tar,	par, car, tar
pole, coal, toll	pole, coal, toll
peas, keys, tease	peas, keys, tease
pan, can, tan	pan, can, tan
Paul, call, tall	Paul, call, tall
pen, Ken, ten	pen, Ken, ten
pool, cool, tool	pool, cool, tool
pill, kill, till	pill, kill, till
pave, cave, tave	pave, cave, tave
puff, cuff, tough	puff, cuff, tough
par, car, tar	par, car, tar
pole, coal, toll	pole, coal, toll
peas, keys, tease	peas, keys, tease
pan, can, tan	pan, can, tan
Paul, call, tall	Paul, call, tall
pen, Ken, ten	pen, Ken, ten
pool, cool, tool	pool, cool, tool
pill, kill, till	pill, kill, till
pave, cave, tave	pave, cave, tave
puff, cuff, tough	puff, cuff, tough
par, car, tar	par, car, tar

SESSION 3, 12

a)	pool, pool, tool	call, call, Paul	till, kill, till
	Ken, pen, Ken	pill, kill, pill	Paul, tall, Paul
	tall, Paul, Paul	tall, call, tall	cool, pool, pool
	kill, till, till	kill, pill, pill	ten, ten, Ken
	pool, cool, pool	cool, tool, tool	call, tall tall
	Ken, ten, ten	pen, ten, pen	pill, pill, kill
	cool, tool, cool	Paul, call, call	Ken, Ken, pen
	ten, pen, pen	pill, pill, kill	tool, pool, pool
b)	pool, tool, cool	pool, tool, cool	pool, tool, cool
	pave, tave, cave	pave, tave, cave	pave, tave, cave
	Paul, tall, call	Paul, tall, call	Paul, tall, call
	pill, till, kill	pill, till, kill	pill, till, kill
c)	pool, tool, cool	pool, tool, cool	pool, tool, cool
	pave, tave, cave	pave, tave, cave	pave, tave, cave
	Paul, tall, call	Paul, tall, call	Paul, tall, call
	pill, till, kill	pill, till, kill	pill, till, kill

SESSIONS 6, 15

a)	tool, pool, pool	Paul, call, call	till, till, kill
	Ken, Ken, pen	pill, pill, kill	Paul, Paul, tall
	Paul, tall, Paul	tall, tall, call	pool, cool, pool
	till, kill, till	pill, kill, pill	Ken, ten, ten
	pool, pool, cool	tool, cool, tool	tall, call, tall
	ten, Ken, ten	pen, pen, ten	kill, pill, pill
	cool, cool, tool	call, Paul, call	pen, Ken, Ken
	pen, ten, pen	kill, pill, pill	pool, tool, pool
b)	pool, tool, cool	pool, tool, cool	pool, tool, cool
	pave, tave, cave	pave, tave, cave	pave, tave, cave
	Paul, tall, call	Paul, tall, call	Paul, tall, call
	pill, till, kill	pill, till, kill	pill, till, kill

SESSIONS 6, 15

c)	pool, tool, cool	pool, tool, cool	pool, tool, cool
	pave, tave, cave	pave, tave, cave	pave, tave, cave
	Paul, tall, call	Paul, tall, call	Paul, tall, call
	pill, till, kill	pill, till, kill	pill, till, kill

SESSIONS 9, 18

a)	pool, tool, pool	call, Paul, call	kill, till, till
	pen, Ken, Ken	kill, pill, pill	tall, Paul, Paul
	Paul, Paul, tall	call, tall, tall	pool, pool, cool
	till, till, kill	pill, pill, kill	ten, Ken, ten
	cool, pool, pool	tool, tool, cool	tall, tall, call
	ten, ten, Ken	ten, pen, pen	pill, kill, pill
	tool, cool, cool	call, call, Paul	Ken, pen, Ken
	pen, pen, ten	pill, kill, pill	pool, pool, tool
b)	pool, tool, cool	pool, tool, cool	pool, tool, cool
	pave, tave, cave	pave, tave, cave	pave, tave, cave
	Paul, tall, call	Paul, tall, call	Paul, tall, call
	pill, till, kill	pill, till, kill	pill, till, kill
c)	pool, tool, cool	pool, tool, cool	pool, tool, cool
	pave, tave, cave	pave, tave, cave	pave, tave, cave
	Paul, tall, call	Paul, tall, call	Paul, tall, call
	pill, till, kill	pill, till, kill	pill, till, kill

MATCHED VOWEL TEST

* stimulus word

SESSIONS 1, 2, 11, 20

a) boy/*bill, whore/*here, *pun/pen, *comb/came, *loop/leap, *ten/ton, *ate/oat, *deal/dual, you/*we, fit/*foot, dare/*door, *coy/keel, *raw/your, could/*kid, *you/we, *mare/more, *anneal/annoy, *your/raw, *where/war, down/*dine, find/*found.

b) *foot/fit, *pun/pen, *anneal/annoy, *door/dare, raw/*you, *comb/came, *bill/boy, *found/find, *kid/could, oat/*ate, *ten/ton, *dine/down, your/*raw, dual/*deal, *you/we, *loop/leap, keel/*coy, war/*where, whore/*here, *we/you, *mare/more.

SESSIONS 5, 14

a) ore/*air, *met/mit, *steap/stoop, hick/*hook, ray/*yell, *core/care, moan/*main, *pen/pun, heat/*hoot, *mind/mound, *put/pit, hook/*hick, hoot/*heat, toy/*tile, *stoop/steap, air/*ore, *care/core, *moan/main, *pun/pen, put/*pit.

b) pit/*put, *heat/hoot, core/*care, pen/*pun, *ore/air, *moan/main, *mind/mound, steap/*stoop, *pit/put, toy/*tile, hook/*thick, pun/*pen, *steap/stoop, care/*core, *yell/ray, mit/*met, moan/*main, *hook/hick, *air/ore, heat/*hoot.

SESSIONS 8, 17

a) fit/*foot, *loop/leap, find/*found, *raw/your, *mare/more, we/*you, *pun/pen, *coy/keel, raw/*your, down/*dine, *ate/oat, you/*we, boy/*bill, *anneal/annoy, *where/war, dare/*door, *ten/ton, could/*kid, *deal/dual, whore/*here, *comb/came.

SESSIONS 8,17

b) *foot/fit, pen/*pun, *anneal/annoy, *door/dare, raw/*your, *comb/came, boy/*bill, you/*we, leap/*loop, keel/*coy, *where/war, whore/*here, *we/you, *mare/more, find/*found, *kid/could, oat/*ate, *ten/ton, *dine/down, your/*raw, dual/*deal.

WORD TESTS

SESSION 1, 20

France, New York, California, Germany, England, Pennsylvania, Italy, Cambridge, Los Angeles, Vermont, Sweden, Texas, Alaska, Soviet Union, Poland, Florida, Virginia, Spain, San Francisco, Boston.

SESSION 3, 12

Detroit, Japan, Hawaii, Nevada, Chicago, Montreal, Greece, Portland, Denmark, Switzerland, Great Britain, Argentina, Ohio, Philadelphia, Hungary, Michigan, Mexico, Connecticut, Norway, Paris.

SESSIONS 6, 15

Argentina, Puerto Rico, Czechoslovakia, Rome, Washington, Denver, Toronto, Indiana, Ireland, Luxembourg, Berlin, Georgia, Austria, New Zealand, Philippines, Colorado, Finland, Rumania, London.

SESSIONS 1, 20

headlight, airplane, sunset, northwest, cowboy, armchair, railroad, hardware, blackboard, sidewalk, cupcake, outlaw, shipwreck, doorstep, wildcat, baseball, mousetrap, hothouse, shotgun, backbone.

SESSIONS 2, 11

yardstick, although, jackknife, playmate, doormat, toothbrush, greyhound, cargo, woodchuck, mushroom, padlock, footstool, eardrum, starlight, pinball, hotdog, midway, washboard, sundown, cookbook.

SESSIONS 5, 14

lifeboat, therefore, beehive, duckpond, workshop, schoolboy, grandson, scarecrow, daybreak, iceberg, blackout, housework, farewell, horseshoe, mishap, pancake, outside, vampire, platform, soybean.

SESSIONS 1, 20

fraud, ride, nook, bar, dish, plush, there, hive, strife, cleanse, wheat, hunt, box, fern, are, rub, heap, pants, manage*, death.

SESSIONS 4, 13

shoe, fate, need, bait, moose, start, cloud, nut, hire, them, wish, bought, quart, pick, frog, tan, blush, five, vast.

SESSIONS 7, 16

1. Glue the sheet to the dark blue background.
2. The juice of lemons makes fine punch.
3. The chest was thrown beside the truck.
4. The hum of bees made Jim sleepy.
5. While he spoke, the others took their leave.
6. The girl at the booth sold fifty bonds.
7. Press the pants and sew a button on the vest.
8. The beauty of the view stunned the boy.
9. Two blue herring swam in the sink.
10. Her purse was full of useless trash.
11. Read the verse aloud for pleasure.
12. He was bribed to cause the new motor to fail.
13. Take the winding path to reach the lake.
14. Haste may cause a loss of power.
15. Cold, damp rooms are bad for romance.
16. A true saint is lean but quite human.
17. A pest may be a man or a disease.
18. The ship was torn apart by the reef.

* included by mistake

SESSIONS 7, 16

19. What joy there is in living.
20. Note closely the size of the field.

SESSIONS 8, 17

feet, table, beans, white, neck, rug, cabbage, grey, eyes, bed, potato, yellow, brain, bureau, lettuce, green, mouth, sofa, cauliflower, purple.

SESSIONS 9, 18

a) liver, feet, nose, neck, stomach, eyes, toes, brain, heart, mouth.
b) lamp, table, chair, rug, closet, bed, chest, bureau, stool, sofa.

SESSIONS 10, 19

a) tomato, beans, celery, cabbage, squash, potato, carrots, lettuce, peas, cauliflower.
b) orange, white, blue, grey, pink, yellow, black, green red, purple.

SENTENCE TESTS

SESSIONS 1, 20

1. Food tastes better when you are hungry.
2. The airplane crashed during the storm.
3. While you are sleeping, you may dream.
4. The red balloon floated into the sky.
5. Read the story aloud to the class.
6. The farmer had many chickens and cows.
7. The bedroom was too small for the brothers.
8. The heavy snow storm stopped all the traffic.
9. The mechanic's hands were dirty and greasy.
10. A smiling face is nice to be near.

SESSIONS 2, 11

1. The freight train traveled along the tracks.
2. Four hours of steady work faced the children.
3. The Frenchman was shot when the sun rose.
4. The water was not fit to drink during the storm.
5. Never kill a snake with your bare hands.
6. A pint of beer made the sailors very happy.
7. The cleaning woman washed the wall and floor.
8. The bowl contained three red fish and a snail.
9. The swan dive was far short of perfection.
10. Hoist the load to your left shoulder.

SESSIONS 3, 12

1. Mend the coat before you go outside.
2. Was he driving the car too fast or too slow?
3. The source of the river is the clear spring.
4. The light flashed the message to the tower.
5. He smoked a pipe until it burned his tongue.
6. The hogs were fed chopped corn and garbage.
7. Glue the picture to the album cover.
8. Death marked the end of his brilliant career.
9. The boys were afraid of the humming bees.
10. The lights should be turned off late at night.

SESSIONS 4, 13

1. The gift of speech was denied the young child.
2. The fire lacked enough fuel to produce heat.
3. Have the suit pressed and cleaned before tomorrow.
4. The salesgirl sold one-hundred tickets to the show.
5. The salt breeze came across the bay into the city.
6. Eating too much makes a woman fat.
7. For quick cleaning, buy a hemp rug.
8. Her purse was full of useless trash.
9. Read the poem aloud if you want to understand it.
10. The car had an empty gas tank and stopped.

SESSIONS 5, 14

1. Because the weather was warm, the children played in the yard.
2. Because the weather was warm, they expected a rain storm.
3. Because the weather was warm, nobody was able to do his work.
4. Because the weather was warm, the meat did not taste like it should have.
5. Because the weather was warm, the flower began to appear in the garden.
6. Because the weather was warm, the storm windows were put away.
7. Because the weather was warm, the boys could not go ice skating.
8. Because the weather was warm, he bought a new automobile.
9. Because the weather was warm, unexpected guests arrived for dinner.
10. Because the weather was warm, the newborn kittens went outside.

SESSIONS 6, 15

1. During the month of May, the soldiers marched into the field.
2. During the month of May, swimming is very enjoyable.
3. During the month of May, the skiing resorts are closed.
4. During the month of May, students take their final exams.
5. During the month of May, the family rented a fishing boat.
6. During the month of May, nobody takes their vacation in Florida.
7. During the month of May, many companies close for two weeks.
8. During the month of May, there is more daylight and less darkness.
9. During the month of May, girls think about meeting boys.
10. During the month of May, the weather becomes very hot.

SESSIONS 7, 16

1. Because school was closed, the instructors did more work.
2. Because school was closed, the coke machine remained full.
3. Because school was closed, the city was almost deserted.
4. Because school was closed, the movie theaters showed cartoons.
5. Because school was closed, too many people went to the beach.
6. Because school was closed, there was less noise in the laboratory.
7. Because school was closed, playing baseball was exciting.
8. Because school was closed, the janitor slept all night.

SESSIONS 7, 16

9. Because school was closed, tickets to the opera were scarce.
10. Because school was closed, newspapers were hard to find.

SESSIONS 8, 17

1. Are you going home this weekend?
2. No, I have too much homework to do for my courses.
3. Forget about the work, you can do it next week.
4. Unfortunately, one of the courses is giving an exam on Monday.
5. Don't you find the exams very easy.
6. This one will be difficult since I have not studied.
7. Do you like taking so many courses?
8. Yes, but I like to go out with girls more.
9. Sometimes there is not enough time for work and pleasure.
10. We all do what we enjoy most.

SESSIONS 9, 18

1. Are you interested in buying a convertible?
2. No, I think a black sedan would be better.
3. The convertible is a very beautiful automobile.
4. But my wife wants us to have a family car.
5. If you buy the convertible, I will reduce the price by 20%.
6. It is not a question of money, I am very rich.
7. The sedan has a maximum speed of 120 miles per hour.
8. But is it a reliable car which I can trust.
9. All our cars have been engineered for reliability.
10. Our last car was always in the garage.

SESSIONS 10, 19

1. My wife and I would like to go on a vacation.
2. I have a special two week trip to Europe.
3. We do not think we can afford to go to Europe.
4. Perhaps you would like to spend three weeks in San Francisco.
5. Yes, that would be very exciting and our parents live there.

SESSIONS 10, 19

6. I recommend that you take a train or bus and see other cities too.
7. How long is the train or bus trip?
8. The non-stop traveling time is about 78 hours.
9. If we go by airplane how long would it take?
10. The direct flight is four hours and you get dinner on the plane.

APPENDIX B

TECHNICAL DESCRIPTION

A block diagram of the electronic system used to implement spectral rotation is shown in Fig. B.1a. The incoming microphone-level signal is first amplified to a more reasonable level and then low-pass filtered to remove any components above 3.2 kHz. Since these components would appear as low frequency energy after being modulated, they must be removed by the filter. The amplifier following the equalizer outputs two phase inverted currents which drive the signal-input side of the balanced modulator. The modulator is simply two electronic switches being opened and closed by the two square wave signals of the 3.2 kHz multivibrator. The effect of the modulator on the spectrum of the input signal is clearly seen in part (C) of Fig. B.1b; the spectrum is shifted by an amcunt equal to the multivibrator carrier frequency.

The second low-pass filter removes the upper-side bands, leaving a single-side band signal with the rather unusually low carrier frequency of 3.2 kHz. This single-side band signal has the property that the output spectrum is the same as the input spectrum rotated about a frequency equal to one-half the carrier frequency.

Technically, there are two major difficulties: 1) the modulator must be balanced to within .5% to guarantee that the leakage of the original signal is less than -45 dB with respect to the rotated signal, and 2) the filters must be sharp enough so that they removed the upper side-band without attenuating the lower one. The first problem is easily solved by realizing that changing the duty cycle of the multivibrator is equivalent to changing the balance. A small adjustment on the multivibrator is provided for this purpose. The second problem is solved by using computer designed elliptic filters with skirts of as much as 300 dB/octave (Saal 1961). Because the filters are chosen for their frequency characteristics, the transient response is not

exceptional good. Typically, the ones used in this system have an envelop delay of about .5 msec., an envelop rise-time of .5 msec., and an envelop overshoot of 2-1/2 dB.

The actual circuits are included for reference and are shown in Fig. B.1c through B.1f.

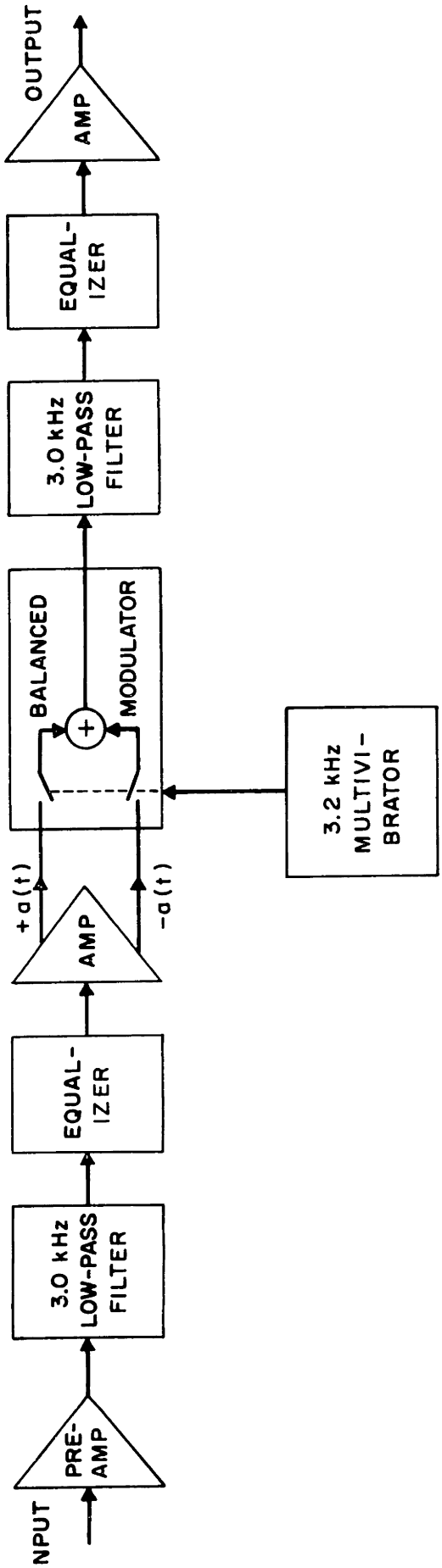


FIG. B.1a. BLOCK DIAGRAM OF SPECTRAL ROTATION SYSTEM.

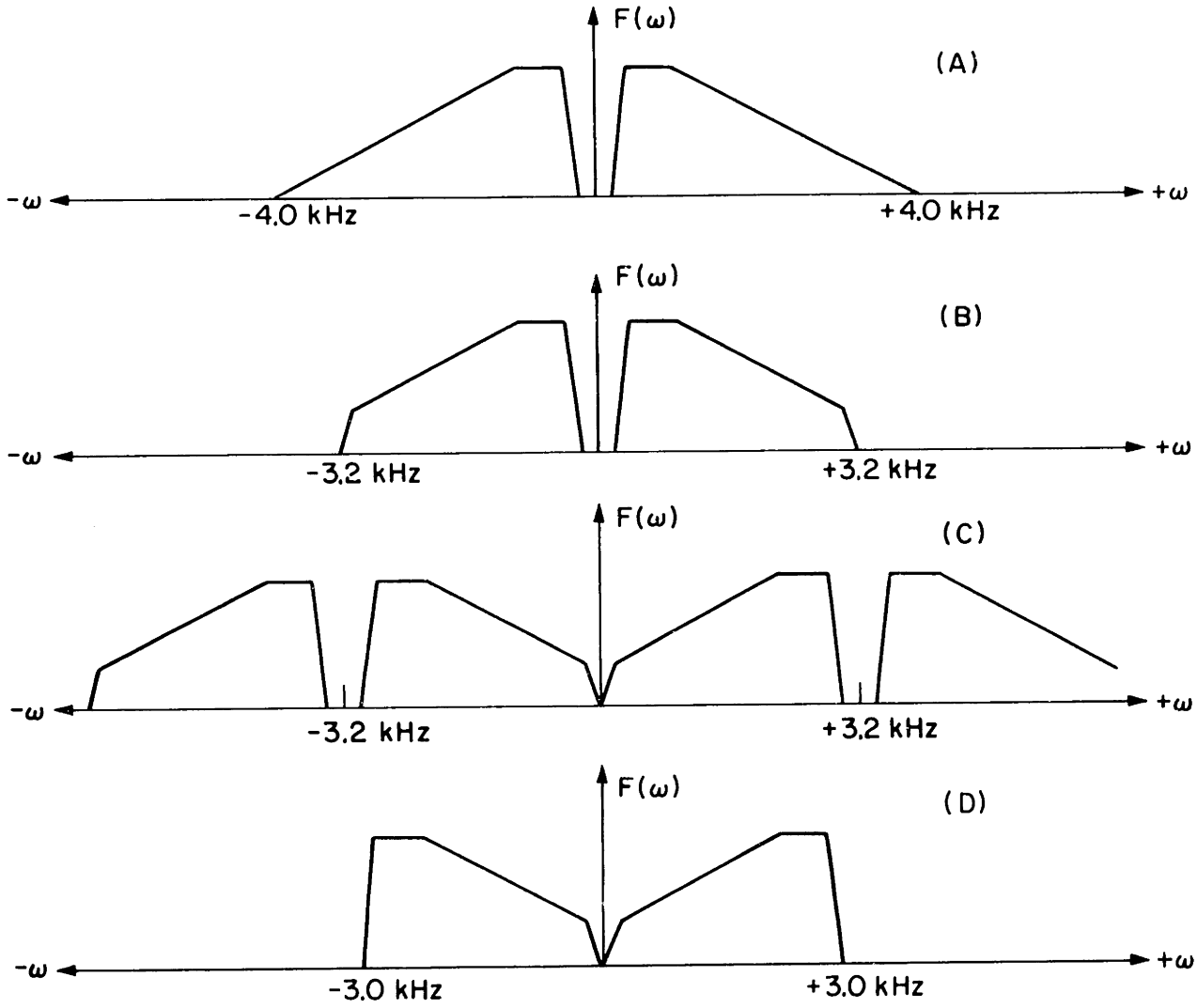


FIG. B.1b. SPECTRA OF SIGNALS IN TRANSFORMATION SYSTEM.

(A) input signal.
(B) after first low-pass filter.
(C) output of balanced modulator.
(D) output after second low-pass filter.

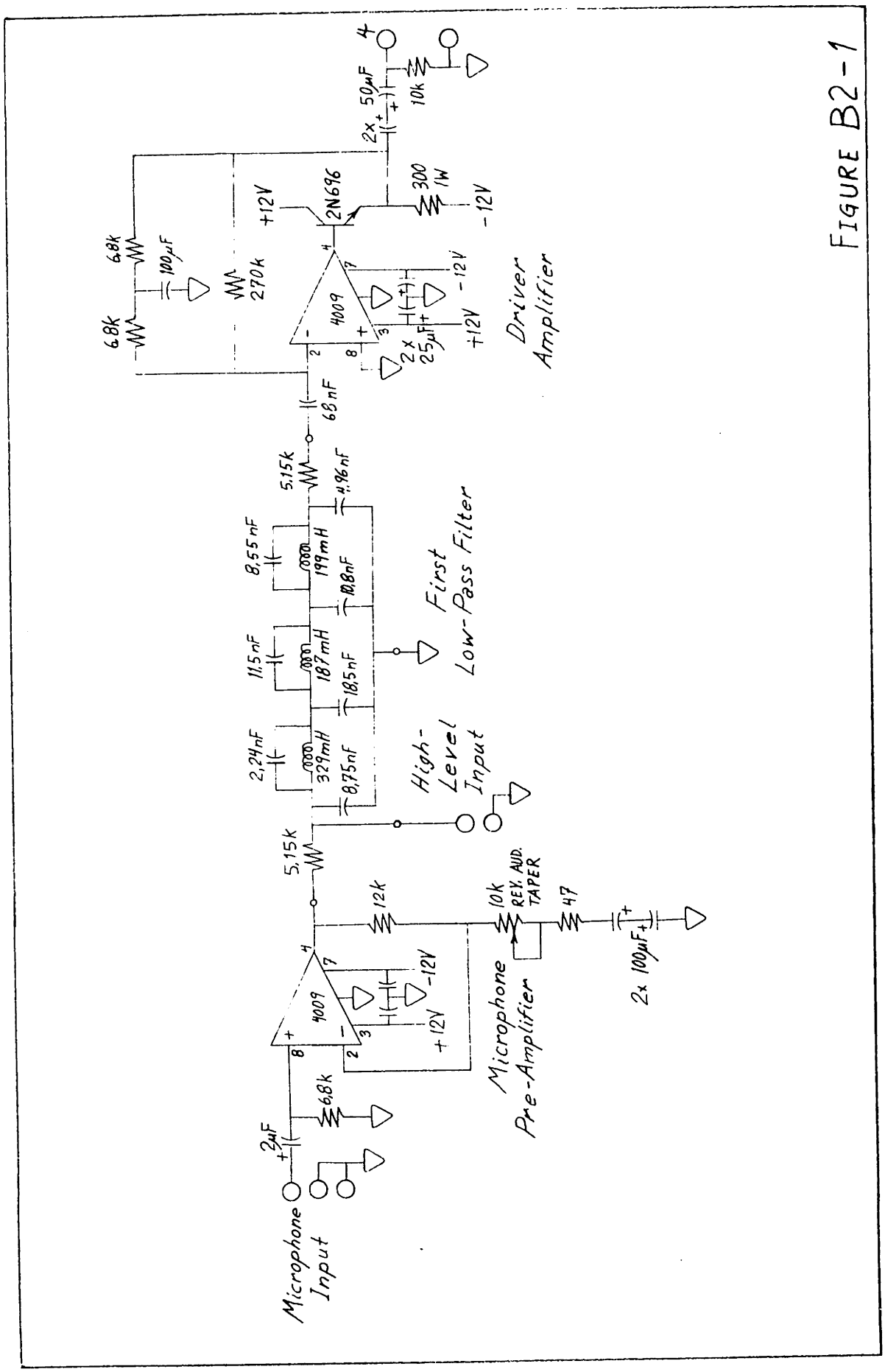
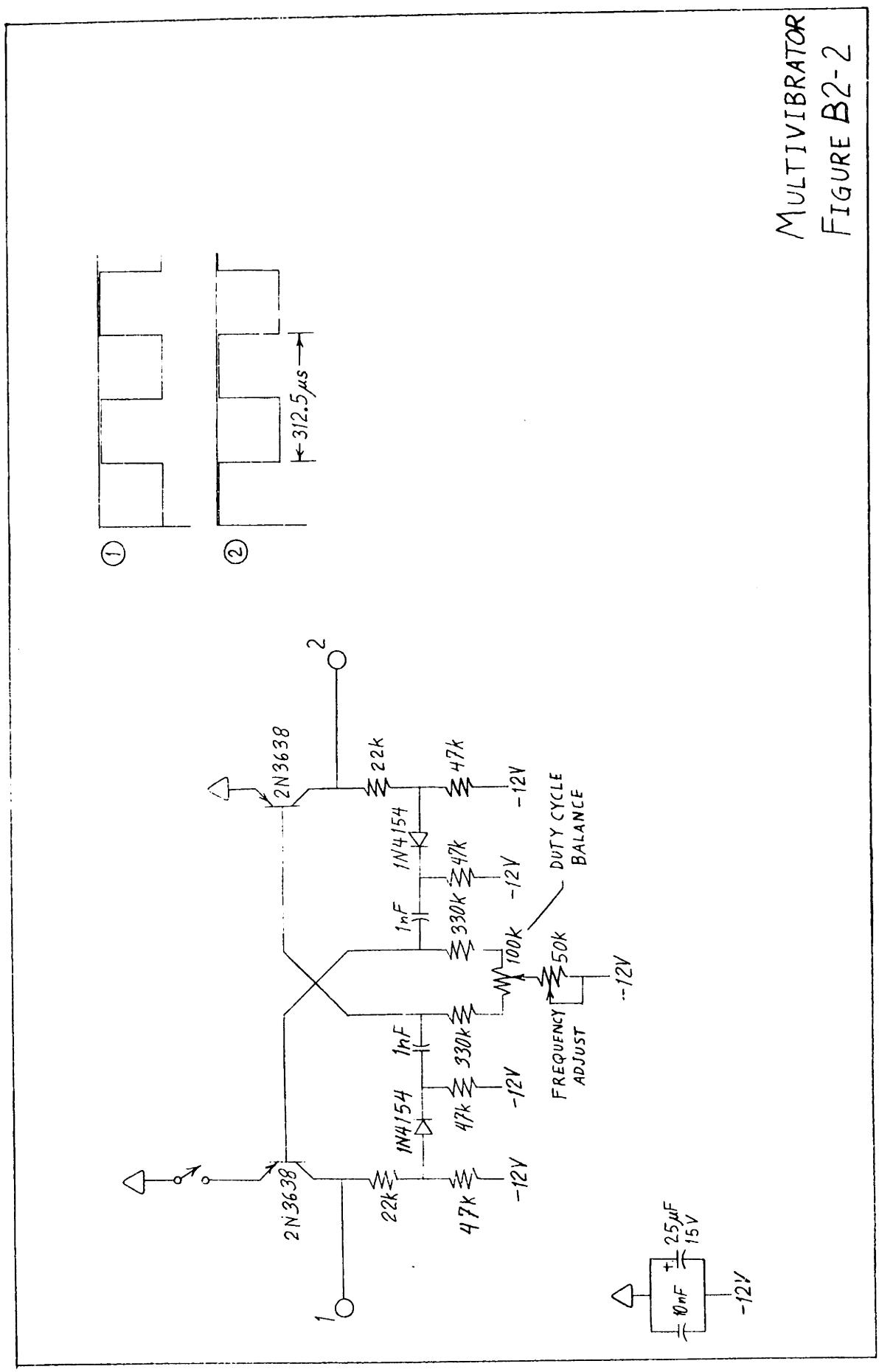
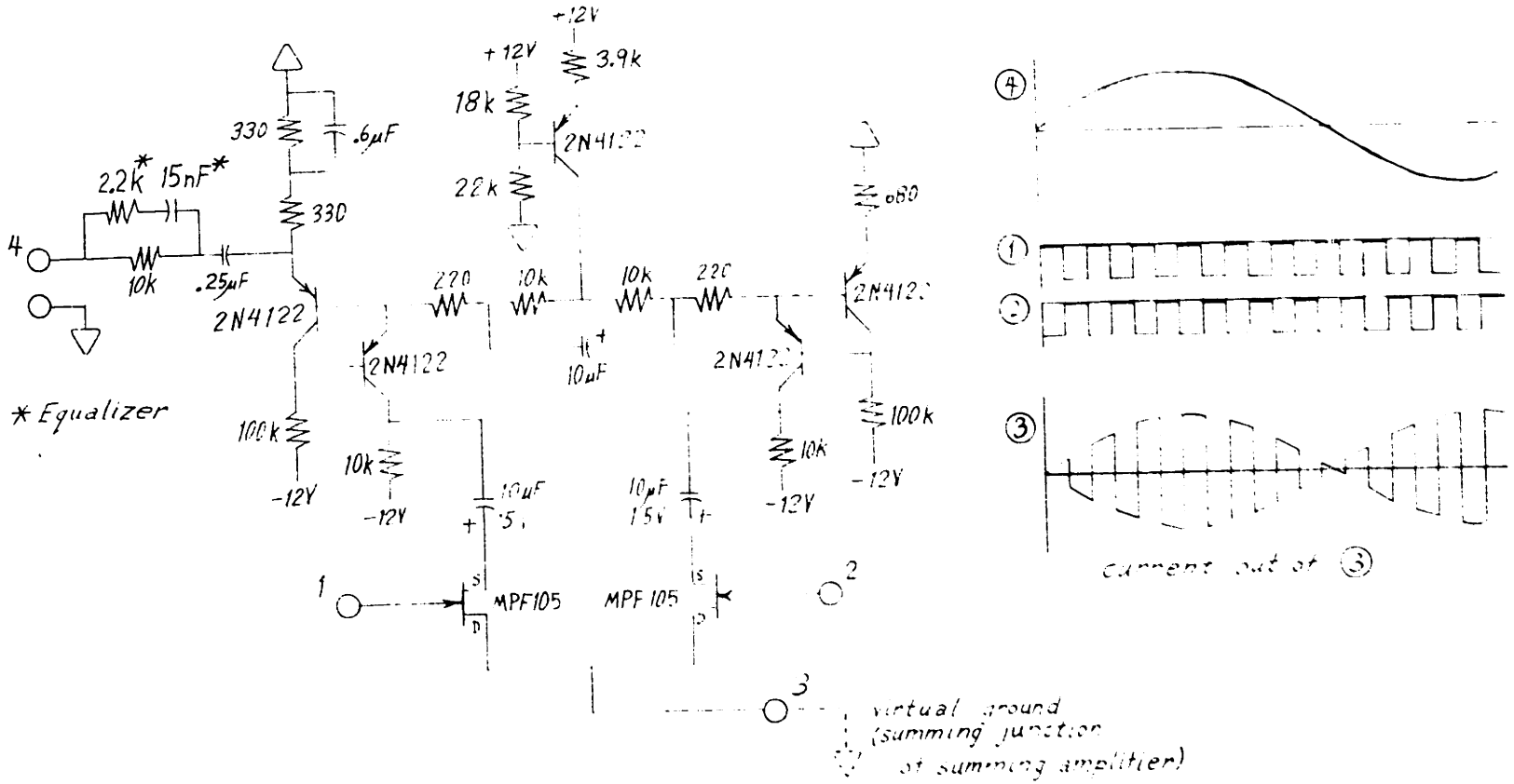

Microphone Input
2μF
6.8k
4009
+12V
-12V
Microphone Pre-Amplifier
12k
10k REV. AUD. TAPER
47
2x 100μF
5.15k
High-Level Input
First Low-Pass Filter
2.24nF
329mH
8.75nF
11.5nF
187mH
18.5nF
8.55nF
199mH
18.8nF
4.96nF
68nF
Driver Amplifier
2N696
300 1W
2x 50μF
10k
2x 25μF
100μF
270k
6.8k

FIGURE B2-1


MULTIVIBRATOR
FIGURE B2-2
①
②
312.5μs
2
1
2N3638
2N3638
22k
47k
-12V
1N4154
1nF
330k
100k
50k
DUTY CYCLE BALANCE
FREQUENCY ADJUST
1N4154
1nF
330k
47k
22k
47k
-12V
10nF
25μF 15V



PHASE
SPLITTING AMPLIFIER
and
CHOPPING MODULATOR
FIGURE B2-3

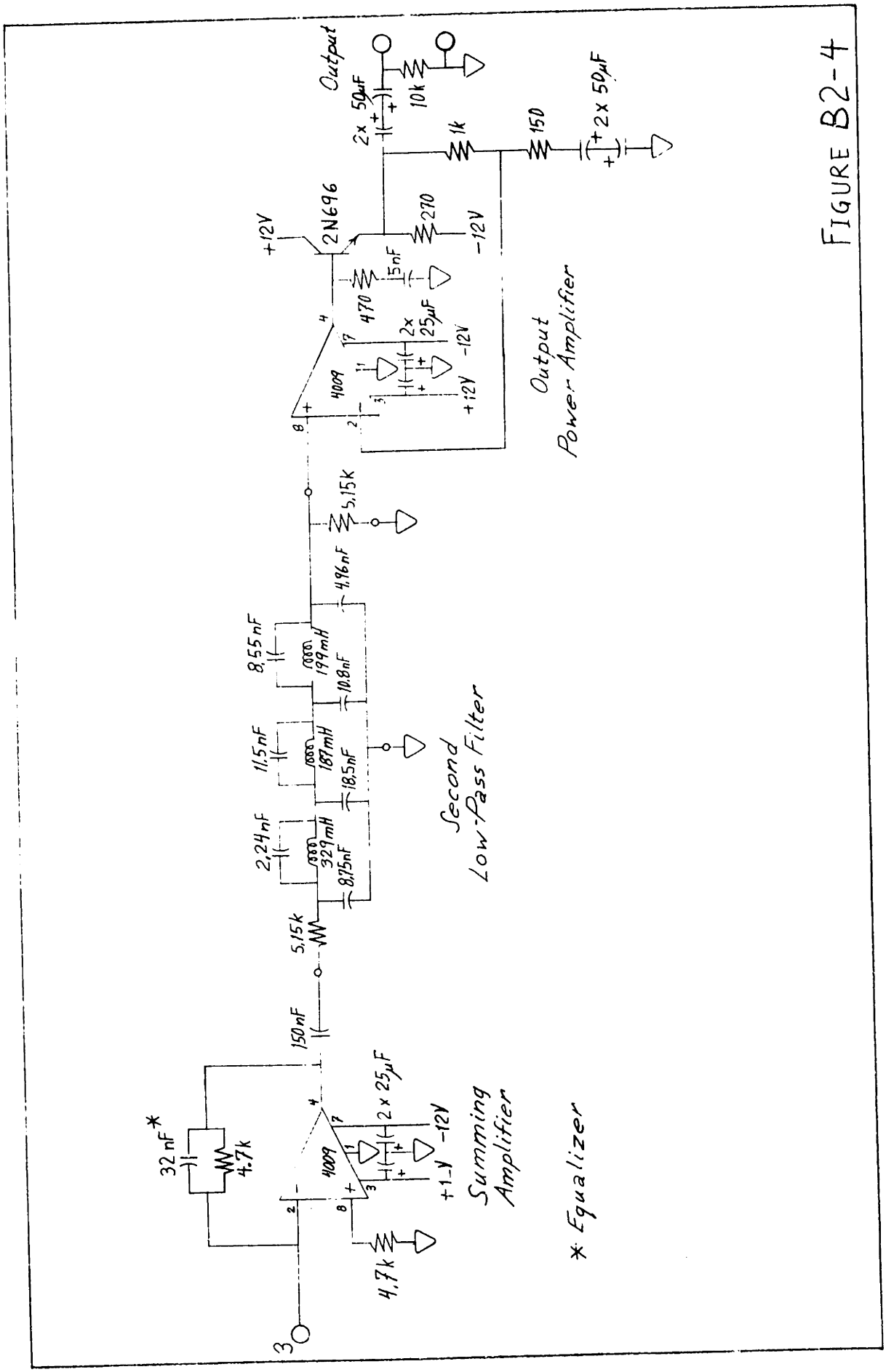

3
32 nF*
4.7k
4.7k
4009
+1.y
-12V
2 x 25µF
150nF
5.15k
2,24nF
329mH
8.75nF
11.5nF
187mH
18.5nF
8.55nF
199mH
10.8nF
4.96nF
5.15k
Second Low-Pass Filter
* Equalizer
Summing Amplifier
+12V
2N696
270
-12V
5nF
470
2x 25µF
4009
+12V
-12V
Output Power Amplifier
Output
2x 50µF
10k
1k
150
2x 50µF

FIGURE B2-4

APPENDIX C

STATISTICAL MEASURE "D"

The statistical measure "D" is used to determine if the kind of practice is a determining factor in learning a particular task. Alternatively, improvement in performance may be the result of exposure to the transformed speech medium and may be determined by the subjects' cognitive structure.

At the beginning of the experiment, the subjects were not matched for any characteristics so that one would expect the pairing to be random. If the nature of the conversation sessions is the important factor in learning transformed speech, then one would expect that the pair of subjects who engaged in the best kind of practice should have ranks of one and two. Similarly, the pair of subjects who engaged in the worst kind of practice should have ranks of eleven and twelve. In other words, if the practice determined performance, then both members of a pair should have very similar ranks. If, however, the conversations were not an important factor, then one would expect the matching of the ranks to be random. The subject with a rank of one should have an equal probability of having a partner with a rank of one or twelve.

The statistical measure "D" is defined as the sum of the magnitude difference between the rank scores of each pair of subjects*. This statistic is implemented by determining if the "D" for a set of ranks is significantly larger or smaller than chance. The probability density function for "D" is easily generated using a computer algorithm which calculates the "D" for each of the 10,395 possible way of arranging twelve ranks into six pairs. The resulting function** is shown in Table C.1a.

*One could define another measure D^2 which is the sum of the squared difference instead of magnitude difference. The results mentioned in Section 6.5 are approximately the same regardless of which measure is used.

**The calculation assumes no tied ranks.

D	p(D)	sum p(D)
6	.0001	.0001
8	.0010	.0011
10	.0038	.0049
12	.0100	.0149
14	.0215	.0365
16	.0360	.0724
18	.0566	.1290
20	.0797	.2087
22	.0970	.3056
24	.1172	.4228
26	.1224	.5452
28	.1293	.6745
30	.1062	.7807
32	.0924	.8730
34	.0577	.9307
36	.0693	1.0000

TABLE C.1a

Probability density function for statistical measure D.

From the data in this table, one can say, for example, that the probability of a randomly paired set of ranks will have a "D" which is less than 10 is .0011; or, the probability of the "D" being greater than 34 is .0693. The data can also be used to generate second and higher-order statistics which allow the null hypothesis to include two or more sets of tests. Thus, one can ask with what probability will the average "D" for four tests be less than or greater than some amount.

REFERENCES

Anisfeld, M. and Lambert, W. Social and psychological variables in learning Hebrew. Journal of Abnormal and Social Psychology, 63, 1961, pp. 524-529.

Bastian, J., Eimas, P. D., and Liberman, A. M. Identification and discrimination of a phonemic contrast induced by silent interval. Journal of the Acoustical Society of America, 33, 1961, p. 842 (abstract).

Békésy, G. von Ueber die mechanische Frequenzanalyse in der Schnecke vershiedener Tiere. Akustica Zeitschrift, 9, 1944, pp. 3-11.

Békésy, G. von Experiments in Hearing. New York: McGraw-Hill Co., 1960.

Bilodeau, E. and Levy, M. Long-term memory as a function of retention and other conditions of training and recall. Psychological Review, 71, 1964, pp. 27-41.

Blickenstoff, C. Musical talent and foreign language learning ability. Modern Language Journal, 47, 1963, pp. 359-363.

Boomer, D. Hesitation and grammatical encoding. Language and Speech, 8, 1965, pp. 148-158.

Boomer, D. S. and Dittman, A. T. Hesitation pauses and juncture pauses in speech. Language and Speech, 5, 1962, pp. 715-720.

Brady, P. T., House, A. S. and Stevens, K. N. Perception of sounds characterized by rapidly changing resonant frequencies. Journal of the Acoustical Society of America, 33, 1961, pp. 1337-1362.

Bryden, M. P. Ear preference in auditory perception. Journal of Experimental Psychology, 65, 1963, pp. 103-105.

Bumstead, A. Distribution effort in memorizing prose and poetry. American Journal of Psychology, 53, 1940, pp. 423-427.

Bumstead, A. Finding a best method for memorizing. Journal of Educational Psychology, 34, 1943, pp. 110-114.

Calvert, L. A. Notes on the Peace Corp language training program. Modern Language Journal, 47, 1963, pp. 319-323.

Cherry, C. On Human Communications. Cambridge, Massachusetts: M.I.T. Press, 1957.

Chomsky, N. Syntactic Structures. The Hague: Mouton & Co., 1964.

Chomsky, N. Topics in the Theory of Generative Grammar. The Hague: Mouton & Co., 1966.

Chomsky, N. and Halle, M. The Sound Pattern of English. New York: Harper & Row, 1968.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. and Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 24, 1952, pp. 597-606.

Daneš, F. Sentence intonation from a functional point of view. Word, 16, 1960, pp. 34-54.

Delattre, P. The physiological interpretation of sound spectrograms. Publications of the Modern Language Association of America, 66, 1951, pp. 864-875.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America, 27, 1955, pp. 769-773.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. Studia Linguistica, 18, 1964, pp. 104-121.

Delattre, P., Liberman, A. M., Cooper, F. S. and Gerstman, L. J. An experimental study of the acoustic determinants of vowel color; observations on one- and two- formant vowels synthesized from spectrographic patterns. Word, 8, 1952, pp. 195-210.

Denes, P. B. Preliminary investigation of certain aspects of intonation. Language and Speech, 2, 1959, pp. 106-122.

Denes, P. B. On the statistics of spoken English. Journal of the Acoustical Society of America, 35, 1963, pp. 892-904.

Denes, P. B. On the motor theory of speech perception. In Proceedings of the Fifth International Congress of Phonetic Science. (eds. Zwirner and Bethge) Basel: S. Karger, 1964.

Dodson, C. J. Language Teaching and the Bilingual Method. London: Sir Isaac Pitman & Sons Ltd., 1967.

Doughty, J. M. and Garner, W. A. Pitch characteristics of short tones: two kinds of pitch threshold. Journal of Experimental Psychology, 43, 1947, pp. 351-365.

Egan, J. P. Articulation Testing Methods. Laryngoscope, 58, 1948, pp. 955-991.

Ehrman, E. Listening Comprehension: In the teaching of a foreign language. Modern Language Journal, 47, 1963, pp. 18-20.

Fairbanks, G. Experimental Phonetics: Selected Articles. Urbana, Illinois: University of Illinois Press, 1966.

Fairbanks, G. and Miron, M. S. Effects of vocal effort upon the consonant-vowel ratio withing the syllable. Journal of the Acoustical Society of America, 29, 1957, pp. 621-626.

Fano, R. M. Short-time autocorrelation functions and power spectra. Journal of the Acoustical Society of America, 22, 1950, pp. 546-550.

Fant, G. Acoustic Theory of Speech Production. 'S-Gravenhage: Mouton & Co., 1960.

Feldtkeller, R. and Zwicker, E. Das Ohr als Nachrichtenempfaenger. Stuttgard: S. Hirzel Verlag, 1956.

Flanagan, J. L. A difference limen for vowel formant frequency. Journal of the Acoustical Society of America, 27, 1955, pp. 613-617.

Flanagan, J. L. Difference limen for formant amplitude. Journal of Speech and Hearing Disorders, 22, 1957, pp. 205-212.

Flanagan, J. L. Speech Analysis Synthesis and Perception. New York and Berlin: Springer-Verlag, 1965.

Fletcher. H. Loudness, masking and their relation to the hearing process and the problem of noise measurement. Journal of the Acoustical Society of America, 9, 1938, pp. 275-293.

Fletcher, H. Speech and Hearing. New York: Van Nostrand Co., 1953.

Fletcher, H. and Munson, W. A. Loudness, its definition, measurement and calculation. Journal of the Acoustical Society of America, 5, 1933, pp. 82-108.

Fodor, J. and Bever, T. The psychological reality of linguistic segments. Journal of Verbal Learning and Verbal Behavior, 4, 1965, pp. 414-420.

Fourcin, A J. Speech source interference. IEEE Transactions on Audio and Electroacoustics, AU-16, 1968, pp. 65-67.

French, N. R. and Steinberg, J. C. Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America, 19, 1947, pp. 90-119.

Fry, D. B. Duration and intensity as physical correlates of linguistic stress. Journal of the Acoustical Society of America, 27, 1955, pp. 765-768.

Fry, D. b. Experiments in the perception of stress. Language and Speech, 1, 1958, pp. 126-132.

Fry, D. B. The development of the phonological system in the normal and deaf child. In Genesis of Language (eds. F. Smith and G. Miller). Cambridge, Massachusetts: M.I.T. Press, 1966.

Fry, D. B., Abramson, A.S., Eimas, P. D. and Liberman, A. M. The identification and discrimination of synthetic vowels. Language and Speech, 5, 1962, pp. 171-189.

Galambos, R. Neural mechanism in audition. Laryngoscope, 68, 1958, pp. 388.

Garrett, M. private communications, 1968.

Garrett, M., Bever, T. and Fodor, J. The active use of grammar in speech perception. Perception and Psychophysics, 1, 1966, pp. 30-32.

Garrett, M. and Fodor, J. Psychological theories and linguistic constructs. In Verbal Behavior and General Behavior Theory (T. Dixon and D. Horton, eds.), Englewood, New Jersey: Prentice-Hall, Inc., 1968.

Geldard, F. A. The Human Senses. New York: John Wiley & Sons, 1953.

Gerstman, L. J. Classification of self-normalized vowels. IEEE Transactions on Audio and Electroacoustics, AU-16, 1968, pp. 78-80.

Gleason, H. A. An Introduction to Descriptive Linguistics. New York: Holt, Rinehart and Winston, 1955.

Goldstein, M. H., Kiang, N. Y. and Brown, R. M. Responses of the auditory cortex to repetitive acoustic stimuli. Journal of the Acoustical Society of America, March 1959, pp. 356-364.

Hadding-Koch, K. and Studdert-Kennedy, M. An experimental study of some intonation contours. Phonetica, 11, 1964, pp. 175-185.

Haggard, M. P. Perceptual study of English /l/ allophones. Journal of the Acoustical Society of America, 42, 1967, p. 1581 (abstract).

Haggard, M. P. and Mattingly, I. G. A simple program for synthesizing British English. IEEE Transactions on Audio and Electroacoustics, AU-16, 1968, pp. 95-99.

Halle, M. In defense of the number two. Studies Presented to J. Whatmough, The Hague, 1957, pp. 65-72.

Halle, M., Hughes, G. W. and Radley, J.-P.A. Acoustic properties of stop consonants. Journal of the Acoustical Society of America, 29, 1957, pp. 107-116.

Halle, M. and Stevens, K. N. Speech recognition: A model and a program for research. IRE Transactions on Information Theory, IT-8, 1962, pp. 155-159.

Harris, C. S. Perceptual adaptation to inverted, reversed and displaced vision. Psychological Review, 72, 1965, pp. 419-444.

Harris, K. S. Cues for the discrimination of American English Fricatives in spoken syllables. Language and Speech, 1, 1958, pp. 1-7.

Heinz, J. M. and Stevens, K. N. On the properties of voiceless fricative consonants. Journal of the Acoustical Society of America, 33, 1961, pp. 589-596.

Held. R. Plasticity in sensory-motor systems. Scientific American, 1965, November, pp. 84-94.

Hirsh, I. J. The Measurements of Hearing. New York: McGraw-Hill Co., 1952.

Hirsh, I. J. Audition in relation to perception of speech. In Brain Function III: Speech Language and Communication. Berkley, California: University of California Press, 1966

Holmgren, G. L. Speaker recognition, speech characteristics, speech evaluation, and modification of speech signal-A selected bibliography. IEEE Transactions on Audio and Electroacoustics, AU-14, 1966, pp. 32-39.

House, A. S. and Fairbanks, G. Influence of consonant environment upon the secondary acoustic characteristics of vowels. Journal of the Acoustical Society of America, 25, 1953, pp. 105-113.

House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. Articulation-testing methods: consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America, 1965, 37, pp. 158-166.

Huey, E. B. The Psychology and Pedagogy of Reading. Cambridge, Massachusetts: M.I.T. Press, 1968.

Huggins, A. W. Distortion of the temporal pattern of speech: interuptions and alternations. Journal of the Acoustical Society of America, 35, 1964, pp. 1055-1064.

Jakobson, R. Selected Writings. The Hague: Mouten & Co., 1962.

Jakobson, R., Fant, G. and Halle, M. Preliminaries to Speech Analysis. Cambridge, Massachusetts: M.I.T. Press (sixth edition), 1965.

Johnson, N. F., The psychological reality of phrase-structure rules. Journal of Verbal Learning and Verbal Behavior, 4, 1965, pp. 469-475.

Kahn, D. The Codebreaker: The story of Secret Writing. London: Weindenfeld and Nicolson, 1967.

Kent, G., The influence of syntactic context on interpretation of discourse. Unpublished paper, Institute of Communications Research, University of Illinois, 1963.

Kiang, N. Y. Discharge Patterns of Single Fibers in the Cat's Auditory Nerve, Research Monograph No. 35. Cambridge, Massachusetts: M.I.T. Press, 1965.

Kimura, D. Left-right differences in the perception of melodies. Quarterly Journal of Experimental Psychology, 16, 1964, pp. 355-358.

Koenig, W. A new frequency scale for acoustic measurements. Bell Laboratory Record, 1949, August, pp. 299-301.

Kohler, I. The formation and transformation of the perceptual world. Psychological Issues, 3, 1964, No. 4.

Kolers, P. A. and Eden, M. Recognizing Patterns: Studies in Living and Automatic Systems. Cambridge, Massachusetts: M.I.T. Press, 1968.

Kozhevnikov, V. A. and Chistovich, L. A. (eds.) Speech: Articulation and Perception. Moscow-Leningrad, 1965.

Ladefoged, P. and Broadbent, D. E. Information conveyed by vowels. Journal of the Acoustical Society of America, 29, 1957, pp. 98-104.

Ladefoged, P. and McKinney, N. P. Loudness, sound pressure and subglottal pressure in speech. Journal of the Acoustical Society of America, 35, 1963, pp. 454-460.

Lambert, W. Psychological approaches to the study of language. Modern Language Journal, 47, 1963, pp. 114-121.

Lane, H. The motor theory of speech perception: A critical Review. Psychological Review, 72, 1965, pp. 275-309.

Lathi, B. Signals, Systems and Communications. New York: John Wiley & Sons, 1965.

Lee, B. S. Effects of delayed speech feedback. Journal of the Acoustical Society of America, 22, 1950, pp. 824-826.

Liberman, A. M., Cooper, F. S., Harris, K. S., MacNeilage, P. F. and Studdert-Kennedy, M. Some observations on a model for speech perception. In Proceedings of the AFCRL Symposium on Models for the Perception of Speech and Visual Form, (Wathen-Dunn, ed .) Boston, November 1964. Cambridge, Massachusetts: M.I.T. Press, 1967.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 74, 1967, pp. 431-461.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. Some cues for the distinction between voiced and voiceless stops in the initial position. Language and Speech, 1, 1958, pp. 153-167.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs, 68 (8, No. 379), 1954.

Liberman, A. M., Delattre, P. C., Gerstman, L. J. and Cooper, F. S. Tempo of frequency changes as a cue for distinguishing classes of speech sounds. Journal of Experimental Psychology, 52, 1956, pp. 127-137.

Liberman, A. M., Harris, K. S., Eimas, P., Lisher, L., and Bastian, J. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. Language and Speech, 4, 1961, pp. 175-195.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 54, 1957, pp. 358-368.

Liberman, A. M., Harris, K. S., and MacNeilage, P. F. A motor theory of speech perception. In Proceedings of the Speech Communications Seminar. Vol. 2, Stockholm: Royal Institute of Technology, 1962.

Licklider, J.C.R. The intelligibility of amplitude-dichotomized, time-quantized speech waves. Journal of the Acoustical Society of America, 22, 1950, pp. 820-823.

Licklider, J.C.R. Audio frequency analysis. In Information Theory: Third London Symposium, (C. Cherry, ed.), New York: Academic Press, 1956.

Licklider, J.C.R., Bindra, D. and Pollack, I. The intelligibility of rectangular speech-waves. American Journal of Psychology, 61, 1948, January, pp. 1-20

Licklider, J.C.R. and Pollack, I. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. Journal of the Acoustical Society of America, 20, 1948, pp. 42-51.

Lieberman, P. Some acoustic correlates of word stress in American English. Journal of the Acoustical Society of America, 32, 1960, pp.451-454.

Lieberman, P. Intonation, Perception, and Language. (Research Monograph No. 38) Cambridge, Massachusetts: M.I.T. Press, 1967.

Lindblom, B.E.F. and Studdert-Kennedy, M. On the role of formant transitions in vowel recognition. Journal of the Acoustical Society of America, 42, 1967, pp. 830-843.

Lisker, L. Minimal cues for separating /w,r,l,j/ in intervocalic production. Word, 13, 1957, pp. 257-267.

Malmberg, B. The phonetic basis for syllabic division. Studia Linguistica, 9, 1955, pp. 80-87.

Marks, L. and Miller, G. The role of semantic and syntactic constrainst in the memorization of English sentences. Journal of Verbal Learning and Verbal Behavior, 3, 1964, pp. 1-5.

Mehler, J. Some effects of grammatical transformations on the recall of English sentences. Journal of Verbal Learning and Verbal Behavior, 2, 1963, pp. 346-351.

Menyuk, P. Private communications. 1968.

Miller, G. A. The magic number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63, 1956a, pp. 81-97.

Miller, G. A. The perception of speech. In For Roman Jakobson (M. Halle, ed.) The Hague: Mouten & Co., 1956b.

Miller, G. A. Decision units in the perception of speech. IRE Transactions on Information Theory, IT-8, 1962, pp. 81-83.

Miller, G. A., Heise, G. A., and Licten, W. The intelligibility of speech as a function of the context of the test material. Journal of Experimental Psychology, 41, 1951, pp. 329-335.

Miller, G. and Isard, S. Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior, 2, 1963, pp. 217-228.

Miller, G. A. and Licklider, J.C.R. The intelligibility of interrupted speech. Journal of the Acoustical Society of America, 22, 1950, pp. 167-173.

Miller, G. and McKean, K. A chromatic study of some relations between sentences. Quarterly Journal of Experimental Psychology, 16, 1964, pp. 297-308.

Miller, G. A. and Nicely, P. E. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 27, 1955, pp. 338-352.

Miller, G. A. and Taylor, W. G. The perception of repeated bursts of noise. Journal of the Acoustical Society of America, 20, 1948, pp. 171-182.

Minnesota, State of, Department of Education A Guide for Instruction in Modern Foreign Languages. Saint Paul, 1965.

Neisser, U. Cognitive Psychology. New York: Appleton-Century-Crofts, 1967.

Orr, D. B., Friedman, H. L. and Williams, J. C. Trainability of listening comprehension of speech discourse. Journal of Educational Psychology, 56, 1965, pp. 148-156.

Peterson, G. E. and Barney, H. L. Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 1952, pp. 175-184.

Plomb. R. Pitch of complex tones. Journal of the Acoustical Society of America, 41, 1967, pp. 1526-1533.

Pollack, I. and Pickett, J. Intelligibility of excerpts from fluent speech: auditory vs structural context. Journal of Verbal Learning and Verbal Behavior, 3, 1964, pp. 79-84.

Pollack, I., Rubenstein, H. and Decker, L. Intelligibility of known and unknown message sets. Journal of the Acoustical Society of America, 31, 1959, pp. 273-279.

Potter, R. K., Kopp, G. A. and Kopp, H. G. Visible Speech. New York: Dover Publications, 1966.

Rabiner, L. R. Digital-formant synthesizer for speech synthesis studies. Journal of the Acoustical Society of America, 43, 1968, pp. 822-828.

Ratliff, F. Inhibitory interaction and the detection and enhancement of contours. In Sensory Communication (W.A. Rosenblith, ed.), Cambridge, Massachusetts: M.I.T. Press, 1961.

Richey, J. L. Taken from the instruction sheet used to demonstrate Record 4599. Bell Telephone Laboratories, 1936.

Richey, J. L. Private communications, 1968.

Ringel, R. L. and Steer, M. D. Some effects of tactile and auditory alteration on speech output. Journal of Speech and Hearing Research, 6, 1963, pp. 369-377.

Robbins, A. M. Speech Sound Discrimination Tests by Robbins & Robbins. Magnolia, Massachusetts: Expression Company, 1948.

Rosenblith, W. A. and Stevens, K. N. On the DL for frequency. Journal of the Acoustical Society of America, 25, 1953, pp. 980-985.

Saal, R., Der Entwurf von Filtern mit Hilfe der Kataloges normierter Tiefpasse. Backnang, West Germany: Telefunken G.M.B.H., 1961.

Savin, H. Word-frequency effect and errors in the perception of speech. Journal of the Acoustical Society of America, 35, 1963, pp. 200-206.

Schott, L. Frequency-Time Division Speech Privacy System: Final Report on Project C-66. Prepared for Division 13 Section 3, National Defense Research Committee, Office of Scientific Research and Development, under Contract OEMsr-795; May 29, 1943; Western Electric Co., Inc.

Schroeder, M. R. Vocoders: analysis and synthesis of speech. Proceedings of the Institute of Electrical and Electronics Engineers, 54, 1966, pp. 720-734.

Shankweiler, D. and Studdert-Kennedy, M. Identification of consonants and vowels presented to left and right ears. Quarterly Journal of Experimental Psychology, 19, 1967, pp. 59-63.

Shannon, C. E. Prediction and entropy in printed English. Bell System Technical Journal, 30, 1951, pp. 50-64.

Shower, E. G. and Biddulp, R. Differential pitch sensitivity of the ear. Journal of the Acoustical Society of America, 3, 1931, pp. 275-287.

Siebert, W. M. Stimulus transformations in the peripheral auditory system. In Recognizing Patterns (P. A. Kolers and M. Eden, eds.), Cambridge, Massachusetts: M.I.T. Press, 1968.

Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Co., 1956.

Skinner, B. F. The verbal summator and a method for the study of latent speech. The Journal of Psychology, 2, 1936, pp. 71-107.

Stevens, K. N. Towards a model of speech recognition. Journal of the Acoustical Society of America, 32, 1960, pp. 47-55.

Stevens, K. N. On the relation between speech movements and speech perception. Proceedings of the XVIII International Congress of Psychology, Moscow, 1966.

Stevens, K. N. The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communications: A Unified View. (David and Denes, eds.) 1967.

Stevens, K. N. and Halle, M. Remarks on analysis by synthesis and distinctive features. In Models for the Perception of Speech and Visual Form. (Wathen-Dunn, ed.) Cambridge, Massachusetts: M.I.T. Press, 1967.

Stevens, K. N. and House, A. S. Speech perception. In Foundations of Modern Auditory Theory (J. Tobias and S. Schubert, eds.) New York: Academic Press, 1966.

Stevens, K. N., House, A. S. and Paul, A. P. Acoustic description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. Journal of the Acoustical Society of America, 40, 1966, pp. 123-132.

Stevens, K. N., Ohman, S.F.G., Studdert-Kennedy, M. and Liberman, A. M. Crosslingual study of vowel discrimination. Journal of the Acoustical Society of America, 36, 1964, p. 1989 (abstract).

Stevens, S. S. Handbook of Experimental Psychology. New York: John Wiley & Sons, 1951.

Stevens, S. S. and Davis, H. Hearing Its Psychology and Physiology. New York: John Wiley & Sons, 1938.

Stevens, S. S. and Volkman, J. The relationship of pitch to frequency: a revised scale. American Journal of Psychology, 53, 1940, pp. 329-353.

Stratton, G. M. Vision without inversion of the retinal image. Psychological Review, 4, 1897, pp. 341-360 and 463-481.

Thomas, I. B. The influence of first and second formants on the intelligibility of clipped speech. Journal of the Audio Engineering Society, 16, 1968, pp. 182-185.

Titone, R. Studies in the Psychology of Second Language Learning. Zurich: PAS-Verlag, 1964.

Tobias, J. V. Relative occurrence of phonemes in American English. Journal of the Acoustical Society of America, 31, 1956, p. 631.

Tonndorf. J. comments from a panel discussion of the Audio Engineering Society "Why do we hear what we hear?" DB, 2, 1968, p. 21.

Voiers, W. D. The present state of digital vocoder technique: A diagnostic evaluation. IEEE Transactions on Audio and Electroacoustics, AU-16, 1968, pp. 275-279.

Warren, R. M. Illusory changes in repeated words: Differences between young adults and the aged. American Journal of Psychology, 74, 1961, pp. 506-516.

Wegel, R. L. and Lane, C. E., The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. Physical Review, 23, 1924, pp. 266-285.

Weir, R. H. Language in the Crib. The Hague: Mouton Co., 1962.

Weir, R. H. Questions on the learning of phonology. In The Genesis of Language. (Miller and Smith, eds.) Cambridge, Massachusetts: M.I.T. Press, 1966.

Wever, E. G. Theory of Hearing. New York: John Wiley & Sons, 1949.

Wever, E. G. and Lawrence, M. Physiological Acoustics. Princeton: Princeton University Press, 1954.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English vowels. Journal of the Acoustical Society of America, 38, 1965, pp. 583-588.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English Consonants. Journal of the Acoustical Society of America, 39, 1966, pp. 388-398.

Williams, C. E., Hecker, M.H.L., Stevens, K. N., and Woods, B. Intelligibility Test Methods and Procedures for Evaluation of Speech Communication Systems. Bedford, Massachusetts: Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, USAF, Hanscom Field, 1966.

Woodworth, R. and Schlosberg, H. Experimental Psychology. New York: Holt, Rinehart and Winston, 1961.

Ziph, G. K. Human Behavior and Principle of Least Effort. Cambridge, Massachusetts: Addison-Wesley Press Inc., 1949.

Zwicker, E. Ueber psychologishe und methodische Grundlagen der Lautheit. Akustica Beiheft (Acustica Zeit Schrift), 8, 1958, pp. 237-258.

Zwicker, E., Flottorp, G. and Stevens, S. S. The critical bandwidth in loudness summation. Journal of the Acoustical Society of America, 29, 1957, pp. 548-557.

BIOGRAPHY

Barry Blesser was born in Brooklyn, New York on 3 April 1943. He attended public schools in Brooklyn and graduated from the Brooklyn Technical High School in June 1960.

He received his S.B. and S.M. degrees in Electrical Engineering from the Massachusetts Institute of Technology in 1964 and 1965 respectively. As an undergraduate, he participated in the co-operative program with the Raytheon Company's Advanced Development Laboratory in Wayland, Massachusetts where he did his research for his Masters Degree on a "Radar Video Data Processor for Air Traffic Control". While he was Technical Manager of the M.I.T. radio station, WTBS, he became interested in audio frequency communication systems with an emphasis on the psychology of perception. This later led to the development of a dynamic range compression system which is now being manufactured by Elektromesstechnik Wilhelm Franz KG, Lahr, West Germany. Several of the concepts used in this system has been published in the IEEE Transactions on Audio and Electroacoustics, Radio Mentor (Germany) and a pre-print of the 35th Convention of the Audio Engineering Society. During the last few years, he has acted as a consultant for special problems in audio frequency systems.

As a graduate student, he had a National Science Foundation Fellowship for one year and, later, a teaching assistantship. Since 1966, he has been an Instructor of Electrical Engineering, teaching a project laboratory on audio communications, and in 1967, he received the Departmental Teaching award of the Supervisors Investors Services Inc. As a member of the Cognitive Information Processing Group of the Research Laboratory for Electronics at M.I.T., he has been engaged in problems of perception of speech.

He is a member of Tau Beta Pi, Eta Kappa Nu, and the IEEE.