A PROGRAM OF RESEARCH DIRECTED TOWARD THE EFFICIENT AND ACCURATE MACHINE RECOGNITION OF HUMAN SPEECH

A Theory of Speech Perception

Final Report No. 2

prepared for

National Aeronautics and Space Administration
Electronics Research Center
Cambridge, Massachusetts

Contract NAS 12-129

Hüseyin Yilmaz
Program Director

November 14, 1967

C-68366

PREFACE

This is a direct continuation of a previous report of the same title* in which a new theory of speech perception was suggested. The theory was built around evolutionary adaptive postulates and applied to whispered speech in general and to vowel perception in particular.

In the present work, we apply the same general approach to consonants. Specifically, we shall consider the consonants *p*, *k* , *č* (cheese), *t* , *b* , *g* , (gone), *ǰ* (juice), *d;* *w*, ɣ (like gh; unknown in English), *y*, and ð (this). Some consideration will also be given to the nasals *m*, *ŋ* (lung), *ɲ* (French ligne); and *n*. We shall not elaborate on *f*, x (Khan; unknown in English), ʃ (shade), *s* ; and *v*, *h*, *z* (azure), *ž*, which appear to be similarly organized. The liquids, *l* and *r*, are probably closer in nature to vowels than to consonants; we shall not consider them in this paper.

The theory led us to a number of predictions which were subsequently tested with the aid of an experimental computer playback system designed for the purpose. To date, the experiments appear fully to support the theory.

We should note that most of these predictions were new and unexpected from the point of view of earlier conceptions, while some others contradicted them. In all cases, the predictions stood up to the test.

The author gratefully acknowledges the generous help and advice of Prof. K.N.Stevens of M.I.T., discussions with Dr.A.G.Emslie, and the technical assistance of J.Shao and W.J.Greene of ADL. Thanks are due to Pastoriza Electronics, Inc., for building and maintaining the digital multiplier system, and to the computer staff of Bolt, Beranek and Newman, Inc., for writing the program and providing the computer facility.

*"A Program of Research Directed Toward the Efficient and Accurate Recognition of Human Speech." To be published in *Bull. Math. Biophys.*, Vol. 29, No. 4, December, 1967, as "A Theory of Speech Perception."

TABLE OF CONTENTS

LIST OF FIGURES

A THEORY OF SPEECH PERCEPTION--II*

Hüseyin Yilmaz

Arthur D. Little, Inc., Cambridge, Massachusetts

ABSTRACT

The author's theory of speech perception, as applied to time-dependent speech sounds, leads to many testable predictions. While some of these predictions are consistent with conventional knowledge, others are new and quite unexpected. A few are in contradiction to long accepted experimental results. A computer-aided experimental program, designed to test the theory, wholly supported these predictions. In view of this outcome, it seems desirable to test other predictions of the theory and to reexamine some conventionally accepted views in order to arrive at a more comprehensive theory of speech. The present findings indicate that, apart from categorization, consonants are similar to vowels: they exhibit parallel organizations and transformation properties.

*Supported by National Aeronautics and Space Administration's Electronics Research Center, Cambridge, under Contract NAS-12-129. The first part of the work will appear in *Bull. Math. Biophys., Vol. 29, No. 4, December 1967.*

I. INTRODUCTION

The everyday uses of vision and hearing appear so unrelated to us that we hardly think of them as analogous or similar to each other. In fact, many obvious differences early led investigators to assume that they must be based on entirely different principles. One of the differences, recognized rather early, was the ear's ability to distinguish two tones from each other when sounded simultaneously. The eye does not distinguish two frequencies. It senses the combination as a new color lying between the two colors. Thus the ear was said to work as an analytical instrument, whereas the eye worked as a synthetic one. This kind of conception puts these two senses poles apart and creates the impression that they have nothing in common with each other.

It is true that there are obvious differences between the two senses, but whether these are differences in principle or only special directions of development within the same general principles, is an important question to answer. In this section, we shall discuss this interesting problem.

Since light and sound are both wave phenomena, we would expect some similarity to exist in the way our perception devices handle them. This expectation is further reinforced when we notice that both color perception and audio perception are object-oriented. In perceiving the objects by their sounds, the ear takes into account the resonant characteristics of these objects. In perceiving the objects by their colors, the eye does essentially the same thing because the absorption characteristics of objects are indeed the resonance properties of the corresponding materials. Of course, any analogy in such widely different areas cannot be expected to hold in all respects but is restricted, instead, to a small range of phenomena. For this reason, we may first delineate some aspects in which no analogy is to be expected. Thus, natural light is phase incoherent, whereas sounds (animal or human) can have phase coherence. Consequently, phase perception in the eye would be out of the question.

Similarly, the ear has a spectral range covering more than eight octaves, whereas the eye covers less than one octave. As a result, the eye cannot be expected to perceive harmonic relationships, whereas the ear is capable of doing so. Further differences between light and sound exist with regard to polarization, quantization, and the noise level in the environment. Each of these may induce a corresponding difference in the two perceptions. When these differences are discarded, there remains a particular area in which the analogy is actually expected to exist (Yilmaz 1967b). This is the perception of wide-band, noise-like sound stimuli (as in whispered speech) by the ear as compared to the perception of wide-band light stimuli (as in color) by the eye. This analogy prompts us to consider, in the audible range, three response functions similar to the tristimulus functions of color vision (Yilmaz 1962).

Imagine that we produced all kinds of noises represented by these three kinds of functions. This is easily accomplished by taking a very wide-range noise generator and passing the noise through combinations of three filters representing our three functions. For this purpose, an amplifier may be fed by the thermal noise of a resistor through three filters corresponding to our functions. After we have obtained our stimuli in this way, we can present them to the ear singly or in various combinations. The question is how the ear will perceive them.

When this is done, the following remarkable property is observed: these distributions sound like various whispered vowels, and they arrange themselves (much the same way as colors) into a vowel circle (Yilmaz 1967b). To the three functions, there correspond the vowels *u* (food), *a* (car), and *i* (feel), respectively. If we present *u* and *a* simultaneously, the ear does not analyze them but a new vowel, between *u* and *o*, like *U* (hood) is heard. If all possible vowels are sounded simultaneously, we hear a neutral vowel like ɝ (bird), or schwa, ə. We can see, therefore, that in the perception of whispered vowels, the ear operates quite similarly to the color perception of the eye

(Figure 1). The analogy is indeed quite general; it sometimes applies to surprisingly uncommon situations such as contrast phenomena which (as we shall see later) exists in consonants. In color, all hues can be obtained from only two colors. This is the essence of E.H.Land's color projections. In phonetics, the analogue of this is the production of all vowels with only two vowels (Yilmaz 1967b). It is gratifying that even this analogy exists, and to a good approximation a two-vowel filtering (designed to produce the analogue of the Land experiment) preserves the original order and identity of the vowels and phonemes.

The analogy between the vowel-space in audio perception and the color-space in visual perception can be exploited further and some useful applications suggested. If a blind man is seated at a table, will he perceive the existence of a jug or glass on the table? Let us remember the "sound of the sea" we hear when we hold a seashell close to our ears. What we hear is the sound selected by the shell out of the surrounding noise in accordance with its resonance characteristics. Similarly, a jug or a glass selects certain frequency regions of the thermal noise. If held close to the ear, the jug can be perceived with ease. The fact that a blind man is more sensitive to sound than are sighted persons, and more experienced in matters of sound, enables him to perceive the jug at a greater distance. The remarkable abilities of some blind people in finding their way around and avoiding obstacles may be due to this sensitivity. (A blinded man, when he first goes out into the world, is frightened by noise. In the long run, however, he learns to perceive by his ears, and noise becomes his best friend.)

Ambient noise, such as the rustle of the wind, the typical street sounds of the city, the factory's usual rumblings, alters the characteristics of sound distribution in much the same way as a change of illuminant modifies light distribution. A necessity for a perceptual transformation similar to color transformation will then arise, for reasons of invariance (stability of the perceived world). Such transformations have not yet been fully investigated, but it is known that they exist and operate essentially the same way as in color.

Carrying the analogy further, a blind man could produce his own noise (e.g., by means of a small transistor noise generator); this would be the counterpart of a sighted man with a lamp or flashlight. This may be something similar to radar. In any case, the analogy seems to be important, and analogues of transformation and contrast phenomena of color perception should be studied with respect to speech. Among many potential applications of this direction of research is the aid to deaf through visual display of his speech. We have already experimented with our display device in this respect. The results have been encouraging.

The perception of objects by sound seems to be perfected by dolphins who, from the echo of their clicks and whistles, determine the nature, size, and distance of various objects. (In their communicative behavior, dolphins appear to produce vowel-like sounds, but at a high frequency range. When transcribed to the human range, these sound like *u, a, e, i,* and *o.)* Bats are insect hunters by ear. They perceive objects and insects by scanning space with a high frequency audio beam. Scanning is necessary because the ear does not form an image, and high frequency is necessary because resolution of small objects is needed. Assuming that a resolution of 1.5 to ~ 1/3 of a millimeter is needed for distinguishing insects, the frequency range usable for bats would be approximately 23,000 to 100,000 cps. How many response functions can the bat have? Here the bat is using his ear for object perception, like an eye. Pattern resolution arguments show that the bat cannot have too many response functions. Bats indeed emit three frequencies: 25,000, 50,000, and 100,000 cps. Of course they can use these as "color quality" receptors. But with this device, they cannot resolve narrow frequency ranges. In fact, it would perhaps be more correct to say that bats have "vision" and color perception (in sound quality), whereas they may be "deaf" in <u>our</u> sense of hearing.

After these comments on object perception through the concept of "illuminant sound," we return to speech perception. A variable sound distribution is analogous to a variable light distribution. Rapid time variations in sound exist as in speech, but a rapid time variation in light does not occur in nature very often. In nature, the flickering of light in contrast to steady light is essentially nonexistent. We therefore expect that in the eye, rapid time variation is not developed far into organized perception. But variation of color from place to place is part of the everyday scene. Therefore a fruitful analogy is expected to exist between time-variable sound distributions and space-variable light distributions (Cooper, et al., 1951). Variations of light and color in space form the basis of our object perception. Similarly, the variation of sound in time seems to be the basis of speech perception. Here, the difference that space is three-dimensional as against the unidimensionality of time, should not deter us; through stereoscopic hearing, the ear tends to perceive location and space as well.

From this point of view, phonemes and words become similar to light patterns extended in space, i.e., analogous to object perceptions. For example, if one of our three functions mentioned above represents the unvoiced plosive *p*, then a time variation according to strength, smoothness, or sharpness, would represent such whispered sounds as *p, b, v,* and *w*. In like manner, the recognition of a great many of the so-called consonants might be performed via simultaneous time variations of the three functions.

Of course, just as in color, contrast- and adaptation-transformations are also to be considered and studied. For example, without establishing the neutral sound as a reference, many whispered vowels themselves sound like noises. This is especially true when the vowel is held a long time. Establishment of reference here is analogous to color coordinates in color perception. It seems that greeting-phrases such as "hello" or "how are you" serve us to establish a reference frame before more essential talk begins.

It is important to emphasize that in matters concerning speech, to establish the validity of certain regularities is not an easy matter. The speech process involves highly variable transformations and normalizations. A given perception is strongly dependent on the phonetic context, grammar and semantic constraints. Furthermore, the human communication channel involves large fluctuations and noise. To separate all these effects and come to the heart of a given regularity is extremely difficult. This state of affairs is reflected in the slow progress of the field and in the variety of the conflicting data presented in the literature. We are hoping that with a logically consistent theory such as we are proposing, it may be easier to discard irrelevancies and to isolate the underlying regularities.

As emphasized in the previous report, our theory is largely independent of mechanism and free of mechanical detail. It relies on a perceptual patterning (in terms of the information-carrying parameters of the physical stimuli), in view of the necessities of life and of the environment. Thus the theory embodies a set of evolutionary adaptive principles which lead to a theory of speech perception studied in its own right. It does not require, nor does it rule out completely, an analysis-by-synthesis theory or different types of motor theories. It goes beyond these theories, however, and suggests a definite perceptual organization with many testable predictions. It is mainly with the testing of these predictions that the theory will stand or fall.

A motor theory of the kind Liberman _et al_. (1963) advocate is really a very general statement from which hardly any specific predictions can be made. Furthermore, as emphasized by Fant (1967), the evidence that can be brought forth by motor theorists could just as well be interpreted in a framework of perceptual theory. On the other hand, the motor theory cannot easily answer such questions as why vowels should possess a circular arrangement, why *t* must have two peaks in its spectral distribution, why *p* and *č* must be complementary and undergo contrast, and why there are perceptual transformations. We believe it would be quite impossible to produce a motor-theory explanation for the special filtering experiments, e.g., the analogue of the Land experiments (Yilmaz 1967b), from a truly motor-theory approach. (For further evidence against motor theories, we refer to a recent work by R.Jakobson (1966).)

In our theory, reference is made to speech-production by our statement that "Perceptual organizations model environmental realities." This requires a matching between production and perception processes. The sound environment for speech is the product of the vocal apparatus and its dynamics. The vocal apparatus and its dynamics, therefore, weigh heavily in our theory. But there is no necessity that the act of perceiving must be directly coupled to motor activity or its controlling centers in the brain. We do not produce light and color with any of our organs, yet we are able to perceive colors and patterns.

Motor theory seems to account for the categorization of consonants rather convincingly. However, as pointed out by Fant (1967), this can also be interpreted as an increased probability at phoneme boundaries, which would fit quite naturally into a (statistical) perceptual theory. Another claim of motor theories, namely the recognition of certain phonemes in absolute terms, is, we believe, inconsistent with the facts because, far from being absolute, these perceptions show a great degree of relativeness. When isolated, phonemes often do not have definite perceptual attributes. A further deficiency of motor theory is that it does not explain satisfactorily why the consonants *p, k, t,* for example, tend to be categorized, whereas the vowels are perceived continuously.

Our theory does not rule out a motor involvement in at least some parts of the perception of speech. But as a general principle of perception, involving vowels, colors, transformations and contrasts, the potentialities of a motor approach appear to be limited.

In our present work, we are trying to construct a general theory of speech perception from adaptive evolutionary principles. These principles fall under two main categories:

A. Physics of the carrier and of the environment.

B. Evolutionary history and the needs of the organism.

These are evidently very general statements. To be of direct use for our constructional purpose, we must make explicit statements within each category. We shall only consider those statements which have a direct relevance to the present task, namely a theory which embodies vowel perception and the perception of some of the consonants. In the first category, we have the statements:

A1. Sound is the carrier of speech information.

A2. Vocal tract modulation is the means of speech production.

A3. Neural material poses no further restrictions.

In the second category, we include the statements:

B1. Perceptual organization models environment.

B2. Perceptual variables optimize survival.

B3. Percepts remain invariant under steady environmental disturbances.

B4. Perceptions caused by short or ambivalent stimuli tend to be categorized.

The last statement reflects the dynamic nature of life (and perception) in which there is always a pressure to decide. When a situation or stimulus lasts a very short time, or if it is ambivalent, there is a need to react quickly or decisively, because some decision is better than no decision at all. Note, however, that this does not mean we must recognize the stimulus in absolute physical terms. Due to the existence of noise and other external and internal conditions, this is not often possible. But there is a need to perceive the stimulus as one of a limited number of alternatives so that we are, in a statistical sense, disposed to make an identification.

In the case of categorized sensory processes, perceptual behavior is strongly suggestive of a discrete set of states in the higher centers of the nervous system. Consider, for example, the case of the ambivalent Necker cube (Figure 2). Here, two different percepts of a discrete sort are available.

Assuming that what is perceived is a linear combination of the two states, we may write:

$$u = \alpha u_1 + \beta u_2 .$$

Since either one or the other of the two states is available, never both simultaneously, we must interpret the α and β coefficients statistically. Thus in the case of the Necker cube, one has one-half probability for each state. The alternations of the percept from one state to the other and back, is a kind of statistical fluctuation. Categorization of a similar kind exists in the perception of certain movements, figures, in the tonal residue, and in the perception of some consonants. For example, the spectral properties of *m* and *n* are virtually identical, but a synthetic sound such as "ana" will be heard some of the time as "ama." Similarly, if *m* is removed from camp and substituted in place of *n* in chant, it will be heard as *n*, not as *m*. In this case, context increases the probability of its being heard as *n*, although physically they are practically the same, owing to the fact that the nasal cavity is fixed.

The property of categorization in speech processes applies more generally than merely at the phonemic level. For example, if the word "Kyoto" is repeatedly pronounced in a steady manner, the perception will shift some of the time to the word "Tokyo."

II. A THEORY OF PHONEME PERCEPTION

A. DYNAMICS OF THE VOCAL TRACT

As we have already stated, the present work is intended as a logical generalization of the previous work on vowel perception. More specifically, we would like to begin an exploration of time-dependent speech sounds by extending our postulational framework in relation to the time-variable. The time-variable enters the speech domain mainly through the dynamics of the vocal tract. So we shall briefly look at the dynamics of this apparatus.

The vocal tract possesses movable configurations as well as static ones. For example, the nasal cavity is essentially fixed, whereas the mouth and lip cavities are variable by virtue of the tongue and the lip movements. Because they are mechanical elements, their responses always take some time. The fastest are the lip and tongue motions. Motions related to the whole jaw or pharynx and chest take longer times. Lips can produce explosions or bursts which, in duration, are as short as 20 milliseconds. This may be a *p*-burst, but in producing the vowel *u*, the shaping of the lips takes a much longer time. In general, vowels take a longer time to initiate, and they are held longer. The consonants of short duration, such as *p*, *k* , *t* , are usually articulated in reference to or in the context of vowels. Consequently, the vowel background shall in principle influence the production and perception of consonants.

In the articulation of *p*, *k* , *t*, and *č* (cheese), there are considerable specializations suggestive of categorization. For example, it is virtually impossible to produce *č* and *k* simultaneously, or to articulate a sound midway between *t* and *k*. However, *pt*, *kp*, *čp* are producible, although they do not occur in language. (In fact if *pt* is produced intentionally in a word, the listener will identify it as either *p* or *t*.) Furthermore, *kt*, which is not producible, sounds almost the same as *pč*; (*kt* is never produced orally but can be manufactured by mixing). In view of these considerations, it seems desirable to postulate categorization of a perceptual nature, as stated in B4, above, although the properties of the oral tract might have helped to promote such a development. In this connection, it may be worthwhile to note that not all languages have identical categorization. For example, Czech exhibits clearly the *p, k, t,* and *č*, whereas in French and English, the *k-č* distinction is not pronounced. The *č* in English is slightly longer than *p, k* , and *t,* and is often classified as a sibilant. Arabic does not have a pronounced *p*. That the perceptual categorization is at least partly learned can be demonstrated by the fact that a monolingual listener usually categorizes speech sounds of a different language according to the codes of his own language.

Our position, then, is that the *p, k, t,* and *č* sounds will have properties endowed them by the existence of speech space, and they will exhibit perceptual organizations and transformations similar to vowel space and color space (Figure 3).

B. PSYCHOPHYSICAL CONSIDERATIONS

The psychophysical power function suitable for the time variable appears to be the linear function

$$s = At + B \qquad (1)$$

where B is a constant defining the origin of time and A is a scale factor (Yilmaz 1967a). Thus, speech sounds must display two quite different invariance properties; a) invariance under the shift of time origin; b) invariance under the change of time scale. The first is fundamental in general, but is not pertinent at the phonemic level. It simply means that an utterance does not depend on when it is produced. The second property implies that under considerable variation in the speed of speech, the phoneme identifications remain unaltered. Indeed, speech is intelligible at 1/2 and 3/2 times the normal speed. At much higher speeds, perception undergoes severe distortion, and intelligibility suffers.

Note that playing a tape twice the speed of the original recording is not equivalent to the speaker's speaking twice his normal speed. In the latter, frequency composition does not change, and it would be understandable at greater variations of speed.

A consequence of the invariance of intelligibility under speed of reproduction is the fact that the frequency composition of bursts and of vowels must show a relativity effect. For example, if a burst, centered around 1000 Hz, sounds like *k* in association with *u*, the frequency of burst must be raised when *a* is used instead of *u*. The full extent of this relativity is expected to be related to the above consideration, namely the frequency range in *k*, *p*, *t*, *č* is predicted to be a ratio $\frac{3}{2} : \frac{1}{2} \simeq 3$ (see Experiment (j) and Figure 5).

C. CONSONANTS AND SPEECH SPACE

We may now call attention to our earlier hypothesis (Yilmaz 1967b) that speech perception may be considered as a time-dependent pattern recognition process in speech space:

$$S(p,\ t) = a_0\,(t)\,u_0\,(p) + a_1\,(t)\,u_1\,(p) + a_2\,(t)\,u_2\,(p) + \ldots \qquad (2)$$

According to this view, a vowel or a continuant corresponds to constant coefficients. Consonants differ from such steady sounds by their *a*-coefficients being time dependent. For example, the stop consonants correspond to extremely short durations, or bursts of vowels. It follows that consonants in general, and stop consonants in particular, will have properties common with vowel space in their frequency compositions. Furthermore, the perception of short and long stimuli will depend on each other in a manner similar to color interactions and contrasts. These ideas lead to a great many predictions and analogies which will have to be tested. For example, the prediction that short bursts, when associated with vowels (see Section A), ought to sound like certain consonants, can easily be tested by a computer-aided spectral filter system. When this is done, one indeed sees that the following associations are obtainable (when referenced to 3):

u burst → *p*	*a* burst → *k*
ě burst → *č*	*i* burst → *t*

The theory then predicts, from the color-theory analogy, that *p* and *č*, and also *k* and *t*, ought to be complementary and undergo contrast transformations. Again, there ought to be invariance under the overall noise, under the change of overall frequency composition, under pitch and intensity, because (following adaptive principles) speech ought to be intelligible under wide variations of personal characteristics and environmental conditions. Furthermore, in the limiting case of pure frequency bursts, one must expect consonant perceptions (not pure tones followed by vowel) when these are associated with a vowel. Details and variations of such predictions and their experimental tests will be presented in the experimental section.

Note that some of these predictions are entirely new, while a few others are contrary to conventionally accepted views. For example, let us compare our predictions with the results presented in a famous paper by Cooper, Delattre, Liberman, Borst and Gerstman (1952). We have the following contradictions: a) There exists a new, *č*-sound in the list of stop consonants; b) *t* has two peaks in its distributions (to complete the circle); c) the burst centered around 1800 Hz, when associated with *u*, will sound like *č*--not like *p*, as they claim; d) the organization given in their Figure 3 is inadequate, since we must include the neutral bursts and vowels inside the circle. Thus, our theory is in conflict with some of their conclusions. Apart from these, we have new and unexpected predictions: a) When associated with a neutral vowel, the *p*- and *č*-bursts, and also the *k*- and *t*-bursts, will be complementary; b) when a neutral burst is associated with *u*, *a*, *ě* and *i*, the perception of the consonants will tend to exhibit contrast, namely *ču*, *ta*, *pe*, and *ki* will tend to be heard. We shall demonstrate, in the experimental section; that all of these, except possibly the *ki*-component just mentioned, are indeed present in speech perception.

When one considers time variations of the same frequency composition, e.g., the *p*-composition, one has a sequence *p*, *b*, *w*, *u*, namely $p \simeq 20$ msec and strong; $b \simeq 50$ msec and less strong; $w \simeq 150$ msec and soft; $u \simeq 500$ msec and continuous. The same is true for, say, the *t*, *d*, *ɣ*, *i* sequence. Here again we seem to observe a regularity: the variation in speed of playing a recording (mentioned earlier) can here be used to predict that these phonemes ought to stand relative to each other at a ratio of approximately three to one. This prediction is approximately satisfied. There seems to be an extra regularity, however; this is the empirical rule that as we go from *p* to *w*, the burst becomes softer. In other words, there seems to be an inverse relation between the intensity of the burst and its time duration in these sequences. Due to lack of time, this point was not investigated in detail. We intend to explore it at a future time.

The contrast and transformation effects exist also in *b*, *g*, *d*, *ǰ*, and in *w*, ϒ, (g^h), *ž*, *y*, as well, but these are weaker and probably more complicated in form. This is expected from analogy with colors, where indeed such organizations and transformations manifest themselves in terms of smaller chips of color. For extended areas of colored patches, other more complicated but less pronounced effects take over.

We must note, finally, that in the present series of experiments, we have considered mainly the burst-quality and duration. Transitional and other cues are removed to make certain that we are dealing only with spectral results. For this reason, the sounds we produce possess a whisper quality and appear somewhat artificial. The transition and aspiration variables were not investigated in detail, although it was concluded from a few restricted experiments that transitional cue aspects probably contribute additively to consonant perception and undergo parallel transformations in speech space. The voice variable (the voiced burst and the voiced vowel) was introduced at several instances, which added further variations to the original experiments. However, the voiced consonants *b*, *d*, *g*, *ž* were investigated only with regard to whisper quality. We intend to remedy this gap at the earliest opportunity.

III. EXPERIMENTAL SECTION

In this section, some of the predictions of the theory with regard to consonants are compared with experiments. Our experimental program was carried out by a PDP-1 computer. Three specially-built A to D and D to A converters and a Grafacon system were used to generate the desired time-dependent signals (see Appendix A). These signals were then passed through the filters for frequency patterning, and the resulting signal was heard directly from a loudspeaker for perceptual evaluation.

According to our theory, consonants are time-dependent patterns in vowel space. Within consonant classes, therefore, there must exist an organization similar to vowels, except that, because of categorization, there will be a smaller number of discernible consonants in each category. The number of vowels can be virtually unlimited since they are not categorized. The very first experiments were directed toward the testing of vowel-consonant analogies and relationships.

a) Whispered vowels of extremely short duration (30 msec or less) were presented to the ear in isolation. In general these are <u>not</u> perceived as speech sounds but identified as clicks. Clicks derived from different vowels had different click qualities, but it was virtually impossible to find speech quality in them. Thus, extremely short bursts of vowels in isolation do not possess speech attributes. The result was the same when the bursts were derived from voiced vowels instead of whispered ones. Bursts derived from pure frequencies led to essentially the same click perceptions, as expected from the theory.

b) The same clicks just described are perceived as stop consonants when followed by a vowel. Thus, when followed by the neutral vowel *ɝ*, the click derived from *u* is perceived as *p*; the click from *a* is perceived as *k*; the click from *ě* as *č*; and the click from *i* as *t*. We emphasize the important fact that although *p, k, č*, have single peaks in their spectrum, *t* must have two peaks, like the vowel *i* (Figure 4), otherwise the perception will not be a satisfactory *t*. This experiment shows that the above stop consonants are patterned like vowels. The number of distinguishable consonants, however, are less in number than the possible vowels because of the categorization property of the consonants (see Figure 2).

c) When, in the above experiment, the frequency composition of the burst is continuously changed, the listeners seem to identify generally only four different consonants. These are *p, k, č*, and *t*. For example, if a burst representing a spectral composition between *u* and *a* is presented, the listener usually identifies it as either *p* or *k*, but not something in between. However, after sufficiently long acquaintance, one begins to discern perceptually some other identifications, for example, *ts, pč*, although these do not occur in the natural language. Untrained observers, on the other hand, under a forced choice procedure, may identify only *p, k, t*, or *p, č, t*, or *t, k, č*. This experiment indicates the nature of categorization and its dependence on training and procedure. The *p-č* and *k-t* pairs are reminiscent of the grave/acute and compact/diffuse features of a conventional distinctive-features classification (Jakobson et al., 1952).

d) When clicks obtained from voiced vowels were used, the results remained essentially the same as in Experiment (b). Moreover, no difference was perceived when these voiced vowels were produced with different voice fundamentals. These experiments show that apart from categorization, the stop consonants are patterned similar to vowels (or, for that matter, similar to colors) irrespective of the voiced or unvoiced quality of the burst.

e) The circular organization just mentioned implies that, as in vowels or colors, there must exist complementary pairs. It appears that *p* is complementary to *č*, and *t* is complementary to *k*. To demonstrate this, we first produced a mixture of all four consonants. This presumably corresponds to a neutral consonant (center of Figure 4). But it is difficult actually to articulate it by the vocal tract, although it has a definite perceptual quality when synthesized by computer and listened to. We denote it by χ. The subject listens and remembers this quality. Later, when p, k, $\check{c}$, and t are presented pair-wise, the same quality is obtained only by $p + \check{c}$ and $k + t$. Other (noncomplementary) combinations such as $p + k$, $p + t$, deviate noticeably from the χ perception.

f) The neutral stop can be synthesized by adding all the other stops, and it can be analyzed to give other stops. Note that although χ is difficult to produce orally (for this, the consonants p, k , t, $\check{c}$ would have to be pronounced simultaneously), it can be fashioned by subjecting a consonant, say t or k, to certain filtering, followed by an amplification. For an application of such filtering, see Experiment (g).

g) Related to the idea of complementarity, it is found that when in a neutral burst we turn the k intensity down, the burst shifts perceptually (usually in a statistical sense) toward $\check{c}$. When we turn p down, the perception shifts toward $\check{c}$, etc. There is in this experiment an overall intensity decrease which may be compensated for by increasing the overall intensity of the burst. This effect is a consequence of perceptual patterning on the speech circle. It is related to the psychophysical function of loudness. (To see this, notice that percepts depend on the ratios of intensities i_1/i_2 and not on the intensities themselves. Hence, reducing the intensity in the p region is equivalent to increasing it in the $\check{c}$ region.) It is similar to the complementarity properties found in colors and vowels.

h) The concept of complementarity leads naturally to an interesting contrast phenomenon. This is investigated by reversing the situation described in Experiment (b), that is, by using the χ-burst followed by the vowels u, a, e, and i. The result (as expected from our theory) is that the χ-burst followed by u is perceived as *$\check{c}u$;* the χ-burst followed by a is perceived as *ta;* the χ-burst followed by e is perceived as *pe;* and the χ-burst followed by i is perceived as *ki* . In these perceptions, *ta* was most easily distinguishable, whereas *ki* was least, and *pe* and *ču* intermediate. In running speech, however, *ki* was just as easily identified. Thus, when running speech was passed through filters so that *k* is flattened to an ƛ -burst, the *ki* perception was still preserved. (The phenomenon is analogous to color contrast where a white spot presented in a colored surrounding will appear roughly as the complementary of the color of the surrounding.)

i) The pure frequency limit of bursts was investigated. When associated with the neutral vowel ɤ, the pure frequency bursts are perceivable (as predicted by the theory) as stop consonants. With ɤ, the following set is typically satisfactory: $p \rightarrow$ 350 Hz, $k \rightarrow$ 1200 Hz, $\acute{c} \rightarrow$ 2400 Hz, and $t \rightarrow \genfrac{}{}{0pt}{}{4500}{350}$ Hz, which are in agreement with vowels *u*, *a* , *é* , and *i* . Note again that for *t* , one needs two separate frequencies. These results parallel Experiment (g) of the previous paper (Yilmaz 1967b) and are directly predicted by the theory. The perception here is not a tone followed by a vowel (as one would presume) but is a consonant followed by a vowel!

j) Another series of predictions from the theory is that the perception of a stop consonant is relative to the vowel immediately following. For example, the noise burst that is perceived as *k* in *ka* is not perceived as such when combined with *u*; it is perceived as *ču* . Thus the frequency composition of a burst does not determine the consonant uniquely. Consonants undergo transformations depending on the vowel immediately following the burst. By a long series of experiments, the spectral shifts shown in Figure 5 were obtained. Along with the results of Experiments (b), (e), and (h), we conclude that these are essentially the results of interactions within the speech space of our theory.*

*Note that a roughly similar but conceptually different organization was previously presented by Cooper et al. (1952). Their work seems to have omitted *ć* because of a forced choice procedure. Complementarity and contrast ideas were not available to them and thus were not studied. Also, it was reported (we believe erroneously) that a burst at 1800-2000 Hz sounds like *pu* when combined with *u* . The predictions of our theory were at variance with this famous article, and one of the reasons for undertaking our experimental studies was this discrepancy.

The maximum shift here is about three times the lowest value of frequency, in agreement with the conjecture of Section B. Incidentally, the influence of the vowel on the preceding consonant raises an interesting question with regard to perceptual causality. In microscopic physics, such a thing would never happen if the principle of causality is valid. However, perceptual time has a resolution width of about 50 msec, and it does not make a sharp distinction between past and future.

k) A similar set of shifts, but smaller in magnitude, are observed for final stop consonants. The stopgap seems to reduce the vowel dependence. (It seems to us that even the final consonant is not strictly a click but actually followed by a short but less saturated vowel, close to neutral.)

l) All the experiments were repeated with bursts derived from voiced vowels and with bursts derived from pure frequencies. Results were essentially the same. No noticeable difference is detected in *k, t, p,* and *č* perceptions.

m) All the experiments were repeated with voiced vowels instead of whispered ones. Results were again the same, with small variations. No intolerable distortion is detected in any of the experiments, including the contrast and complementarity phenomena. In general, it seemed that the results were easier to produce and interpret if whispered vowels were used instead of voiced ones. As we have claimed all along, organization is more simply manifest in a whisper or whisper-like synthetic speech (Winckel, 1967).

n) Continuous speech was passed through various filters, including the one similar to two-color projections (Yilmaz 1967b). Although, with these filtering actions, some of the bursts were severely distorted (or sometimes virtually eliminated), they were nevertheless perceived in the speech. This experiment attests to the overall relativity of speech perception. Related to such severe transformations, when *p, t, k, č* were shifted simultaneously to higher frequencies, their perceptual order still seemed to be preserved even without shifting the vowels themselves. However, because of lack of time, and of the extremely complex procedures involved, this line of investigation was not fully pursued.

o) When a noise burst, e.g., *p* , is extended in time and reduced in intensity, its perception is shifted to *b, v,* and eventually to *u.* Similarly, *t* shifts to *d,* ȝ , and eventually to *i* (see Figure 6). The results of such experiments are given in Figure 9. We note that a nasal aspiration preceding *b, d, g,* and *ž* helps their identification considerably.

p) The nasals *m, n,* ŋ, and ɲ (French ligne) appear to differ so slightly in their physical aspects that they are ambiguous. Any one of them is easily turned into another simply by putting it into another context (they are highly categorized). When not in a proper context, as in ana, the perception will shift to ama, aŋa, and back to ana, after (sufficiently) hearing them repeatedly. The probability of perceiving a given nasal depends mostly on context and linguistic association rather than spectral properties.

q) The fricatives *f,* χ (Khan), the sibilants ʃ and *s,* and also *v, h, z* , and *ž,* appear to have the same parallel to *p, k* , *č* and *t* . For example, short *s* (with adjustment of burst power) sounds like *t,* whereas short *f* sounds like *p,* etc. These are not investigated in sufficient detail, but if the general organization suggested above is correct, all the phonemes, except possibly *ℓ* and *r,* will appear to be patterned according to our theory.

r) Vowel space, complementarity and contrast effects are demonstrated with a neutral burst followed by continuously varying the vowels. When the vowel varied from *u* through *a* and to *ě*, the perception of the burst moved from *č* through *t* and to *p*. This experiment is quite interesting because one is continuously able to compare the consonants heard.

IV. DISCUSSION

As emphasized in Section I, this paper deals with a generalization of some of the concepts of our perceptual theory to time-dependent speech sounds. In so doing, we believe we have brought forward new evidence as to the essential validity of this theory. We would like to summarize the most striking of the findings for easy reference.

a) Stop consonants *p*, *k*, *č*, and *t* are patterned similar to vowels *u*, *a* , *e* , and *i* on the vowel circle. In particular, *t* has two maxima like *i*. A summation of *p*, *k*, *č*, and *t* leads to a neutral consonant *χ*, which is perceptually present, yet it is hardly ever produced by man. The complementaries *p* and *č*, and also *k* and *t*, produce the same neutral *χ* sensation. The perceptions of *p*, *k* , *č*, and *t* tend to be categorized.

b) Without the association with a vowel, the bursts representing these consonants are perceived as clicks. Furthermore, both in the sense of contrast and in the sense of transformations, the spectral distribution of bursts does not determine uniquely their perceptual properties. To a large extent, there exists a relational relativistic patterning in the perception of these consonants.

c) Continuous speech, when filtered by various filters, including the one similar to two-color projections, remained intelligible, and all consonants as well as vowels were clearly discernible, including *t*. This shows the existence of overall stability and relativistic transformations in speech space.

d) Continuous speech, when speeded up or slowed down within certain limits, preserves its intelligibility, including the identifications of vowels and consonants. This shows that the overall frequency shift in vowels and consonants tends to make no difference, again pointing to a relational, relativistic patterning of phonemes in time and in frequency, in accordance with perceptual and psychophysical laws. This is the condition of the perceptual stability of the external world.

e) A combination of spectral and temporal aspects appears to cover essentially all the aspects of speech sounds and their perception. It is important to remember, however, that perceptual aspects of speech possess and obey laws which are not in one-to-one correspondence with the physical aspects of the sound distributions. Perceptions have their own laws and transformations which are beyond the merely physical aspects of stimuli.

f) It appears that these experiments contradict motor theories of a simpler kind. For example, in a motor theory there would be no way of understanding why an $\mathfrak{r}$-burst followed by $\mathbf{a}$ should sound like ta when an a-burst followed by $\mathfrak{r}$ sounds like $k\mathfrak{r}$. In our theory, this is a consequence of the complementarity of k and t. Motor theories here are helpless because an $\mathfrak{r}$-burst is impossible to articulate. It would necessitate articulating all of p, k, $\check{c}$, and t simultaneously. Even if this were possible, there would still be no reason why k and t should be complementary and undergo contrast transformations. (As in Hering's opponent-representation of color space, the pairs p-$\check{c}$ and k-t represent opposite sensations.)

g) Similarly, the pure frequency limit for p, k, $\check{c}$, and t are not derivable from a motor theory unless, of course, it is postulated that motor commands of the brain are in terms of pure frequencies. But then there would be no explanation why **t** is articulated in terms of two frequencies, whereas others have only one. In our theory, the two frequencies are necessary to complete the mapping on a circle.

A more sophisticated motor or articulatory theory might account for these and other experiments, but our guess is that such a theory will have to be exceedingly complicated and contain many _ad hoc_ assumptions. Our theory incorporates articulatory aspects with the statement, "Perception devices model environmental properties." Since the speech environment is a product of the vocal tract and its motions, it would seem that this carries us, as far as it is needed, functionally. An articulatory theory tends to operate in an absolute sense, and the highly involved transformations of Experiment (j) would be quite difficult to justify, let alone to explain, in such a theory.

We must emphasize, however, that the present theory does not really rule out such ingenious models as the analysis-by-synthesis theory of Stevens and Halle (1967). However, in their present state of development, these models are statements of a very general nature and predict little specific patterning. Furthermore, it would be quite difficult to prove a motor theory because all its consequences can also be accounted for by a perceptual theory. Thus, the motor parts may or may not be actively involved; it makes no difference to a perceptual theory. (In these respects, see the discussion by Fant, 1967.) Besides, if we accept the color-speech analogy, a motor theory of the type of Liberman et al., would not be general; after all, we perceive colors and patterns without actually producing light and patterns ourselves! To quote Fant (1967), "the motor theory of speech ...will shed a new light on the acoustic structure of speech, but we should not ignore perceptual patterning simply by a reference back to production." Still, our main objection to motor theories is not on the basis of their aesthetic quality but rather on their ability to produce crucial testable predictions. The value of a theory lies in its susceptibility to material disproof, since no theory can ever be proven right.

Our theory is quite consistent with the "distinctive features" of Jakobson, Fant and Halle (1952). For example, "grave/acute" and "diffuse/compact" separations correspond to dividing the vowel circle into four quadrants (Figure 8), whereas the "tense/lax" distinctions are related to our saturation considerations, as in *u*, *o*, *U* , ɝ, etc. However, the vocalic/unvocalic distinction is not a feature which is absolutely necessary for speech intelligibility. This is clear from the fact that whispered speech is perfectly intelligible. In general, the "distinctive-features" approach is division-oriented and tends to suppose that these separations are binary (information theory in binary form probably played a role here) and absolute. In comparison, our theory may be said to be element- and categorization-oriented, and it recognizes that even the distinctive features must be relative in their physical attributes (Jakobson, 1966; Hiramatsu et al., 1967). From the distinctive features alone, it is not obvious why *p-č* and *k-t* should be complementary pairs and undergo contrast. Furthermore what distinctive feature must we use for a classification of desaturated vowels like U? Such questions appear to indicate that the "distinctive features" approach can become quite artificial if we insist on simple binary separations. Our theory seems to be capable of incorporating the "distinctive features" in a modified, perhaps slightly more complicated form. From our point of view, the "distinctive features" approach is not a theory but a datum to be explained by another more general theory.

Our theory promises practical applications in the following areas:

a) Automatic Speech Recognition by Computer. This is a highly desirable goal, since it would open the way to great developments in industry and social life by making possible the vision of electronic banking, computerized shopping, speedy long distance transactions and communications, etc.

b) Spoken Command and Control of a Machine by a Human Operator. This is part of the overall subject of man-machine communication, and our theory could help to advance the field by making accurate and efficient recognition of speech easier. This would then lead to computerized design and production methods operated by voice and speech. Already recognition machines to process mail by spoken zip coding are becoming operational, and the spoken dialing of the telephone will probably follow. When hands and feet are occupied by other functions, such as required in the busy schedule of an astronaut in space flight, or a pilot in fighter aircraft, control over machinery by voiced command is a very strong desideratum. Also, when an operator is placed in a noisy environment such as in a helicopter, voice qualities can be displayed to him visually.

c) Programming a Computer by Voice. This goal could be accomplished if 200 to 300 words can become recognizable by machine with sufficient accuracy and without regard to speaker. Then every telephone in every home and office in the country can gain simple access to a computer, and the market will expand immeasureably. Control of a computer by voice is probably possible if only 50 to 60 words are recognizable, irrespective of speaker and ambient noise.

d) Speech Communication at Reduced Cost. The more we learn about the true nature of speech, the more we shall be able to remove redundancy in transmission and to reduce cost.

e) Speech Transposition from Different Speeds and Frequencies. Through our psychophysical considerations and detailed transformations, it follows that speech transcription from, say, underwater speech to normal speech is possible with better efficiency. Also, speeded speech provides much savings in time, e.g., in listening to lectures, etc.; conversely, slowing down speech can be useful in studying a foreign language, with improved efficiency, etc.

f) Visual Aid for the Deaf. Our theory gives specific prescriptions for a nonredundant optimum representation of speech in terms of a color circle on a scope face. This is a highly desirable representation, and deaf subjects appear to prefer it. Representation includes both vowel and consonant information. After the normalizations with respect to various transformations, this could be an efficient way to teach the deaf how to speak. Representation directly in terms of color variables is also possible on a color television tube. In fact, a tickertape kind of representation, including a person's last two seconds' length of speech, could be constructed (Benninghof 1967). This could in principle include context effects by utilizing such transformations that the eye already possesses in terms of color.

APPENDIX A

EXPERIMENTAL APPARATUS

In order to test the usefulness of the speech cone suggested in our proposal, we have designed the three-channel system shown in Figure 7. This system accomplishes the multiplication of spectral functions $u_1(\nu)$ $u_2(\nu)$, and $u_3(\nu)$ by time functions $f_1(t)$, $f_2(t)$, and $f_3(t)$, respectively, and sums the resulting products:

$$f(\nu, t) = u_1(\nu) f_1(t) + u_2(\nu) f_2(t) + u_3(\nu) f_3(t).$$

It is conjectured in our proposal that this function $f(\nu, t)$ represents, to some approximation, the simple speech sounds such as *na*, *da* , *ta* , *ma* , *ba* , *pa* , etc. In our system, $u_1(\nu)$, $u_2(\nu)$, and $u_3(\nu)$ are fixed functions of frequency, and they represent the vowels u_1, u_2, and u_3. The time functions $f_1(t)$, $f_2(t)$, and $f_3(t)$ are to be generated by the computer with the help of curves drawn by hand on a screen. We shall be able to change or modify these functions any way we like and to study their perceptual effects and transformations by listening to the corresponding acoustical distribution generated by the computer.

The functions $f_1(t)$, $f_2(t)$, and $f_3(t)$, once determined for a syllable or word, are stored on tape. These functions can then be generated by a PDP 8 digital computer which is situated at Bolt, Beranek, and Newman, Inc., Cambridge, Massachusetts. There are three synchronized output channels from the digital computer, one for each function. Each channel handles six bits. There is a maximum of 64 combinations available as output. The multiplication processes are accomplished by using digital-to-analogue multiplying units especially designed for this purpose. These digital-to-analogue multiplying units are modified from the commercially available DC digital-to-analogue converters with special circuits incorporated to facilitate AC multiplication.

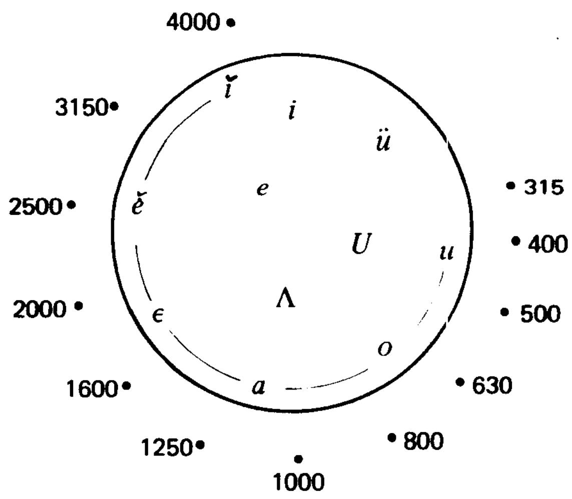


Figure 1. The vowel circle of human speech. Included in the vowel category are *u* (as in the French *rue*); *e* (saturated *ĕ*, which sounds close to *ĭ*); *ĭ* is a saturated *i* with no low frequency component.

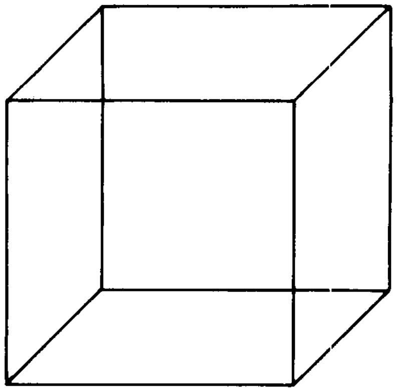


Figure 2. Necker's cube. The perception referring to solid cube is categorized and oscillates back and forth. The cube is seen either in one way or the other, but never both simultaneously.

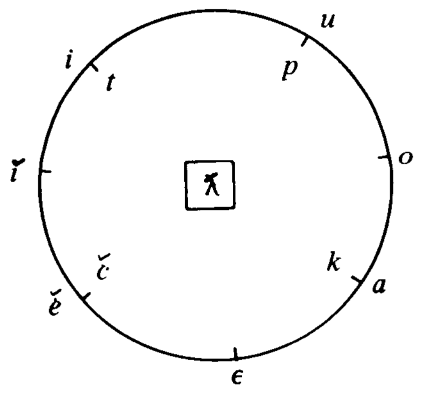


Figure 3. The consonants *p* , *k* , *c* , and *t* can be represented (formally) on a vowel circle due to their spectral similarity to vowels. Nnte again the *u-p* , *k-a* , *č-e* , and *t-i* proximities. The neutral consonant, λ̌ , is not producible by the vocal tract. The *e* and *i* correspond to saturated frequency areas with center frequencies 2400 Hz and 4000 Hz, respectively.

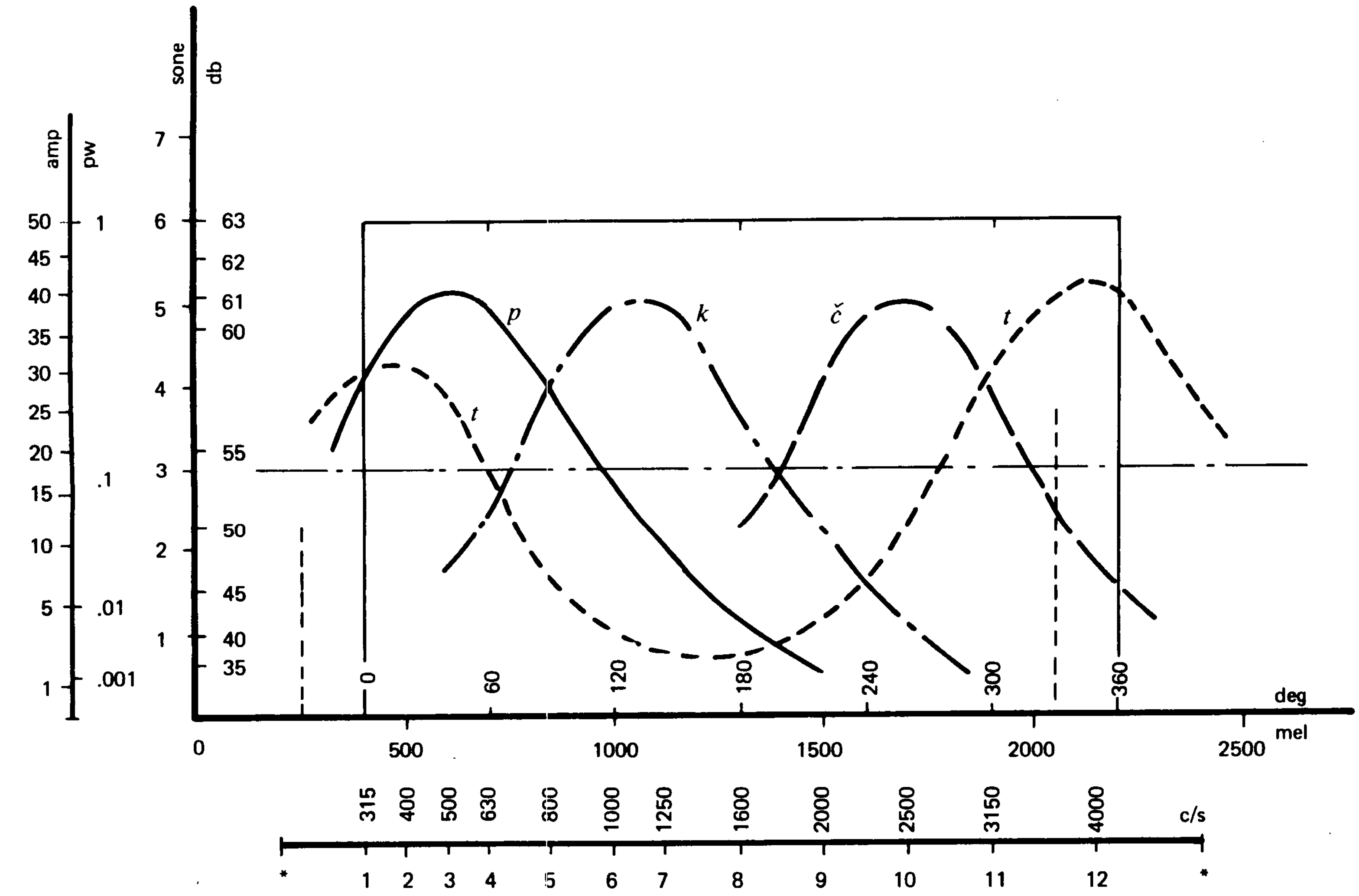


Figure 4. Spectral properties of noise bursts *p* , *k* , *č* , and *t* , when pronounced as in *pa* , *ka* , *ča* , and *ta* . Notice the similarity of these spectra to the spectra of *u* , *a* , *e* , and *i* .

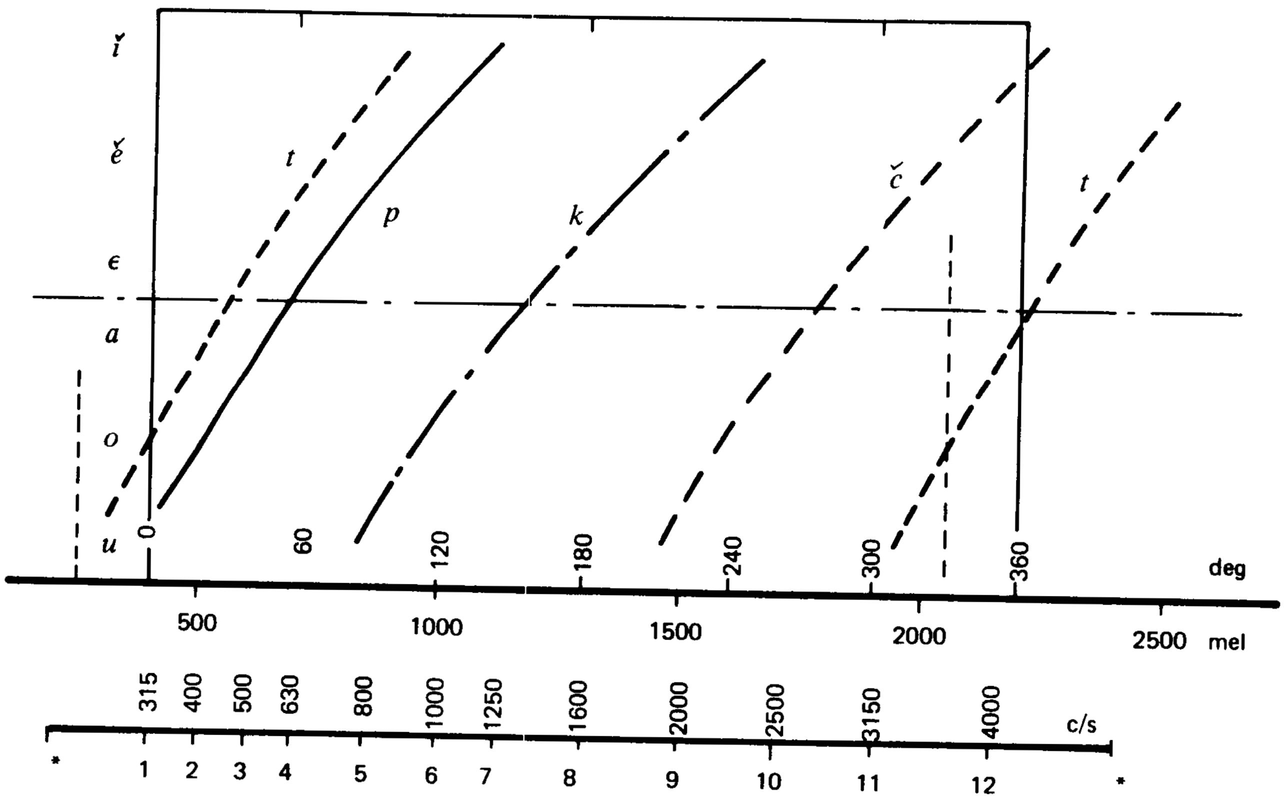


Figure 5. Spectral transformations of *p* , *k* , *č* , and *t* , when pronounced with various vowels. The systematic shift observed is evidence of the relative nature of consonant perception.

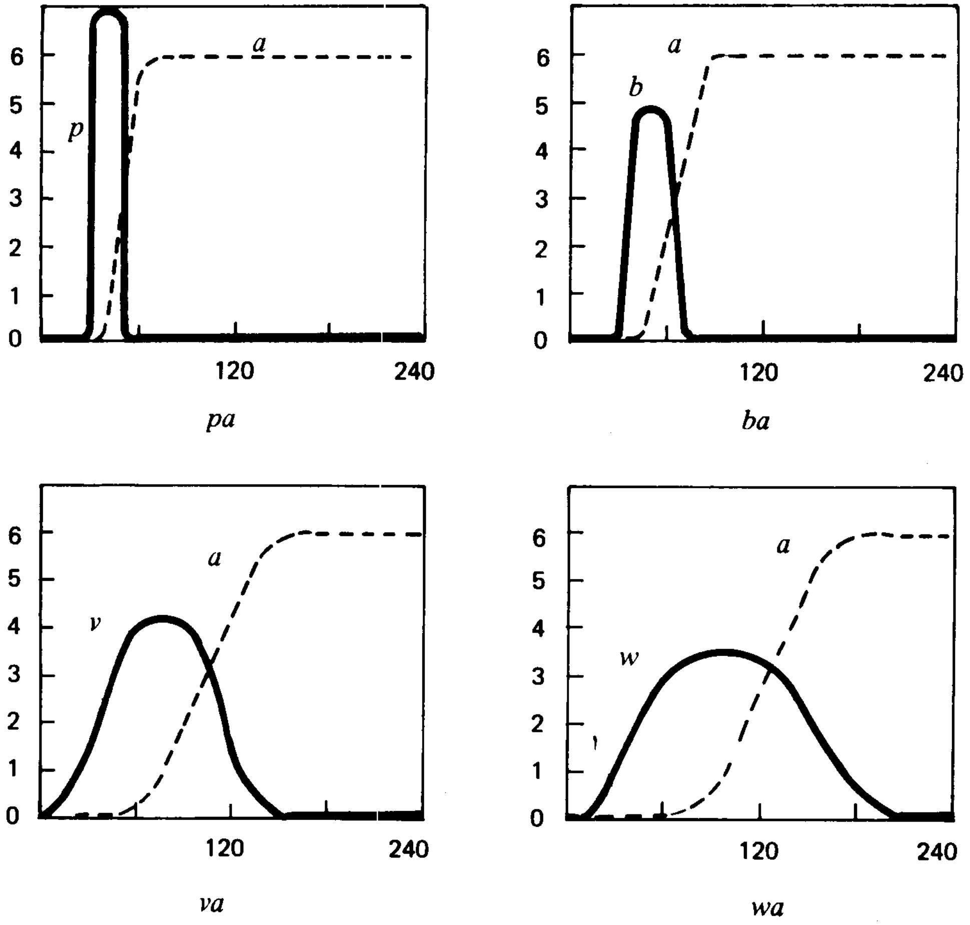
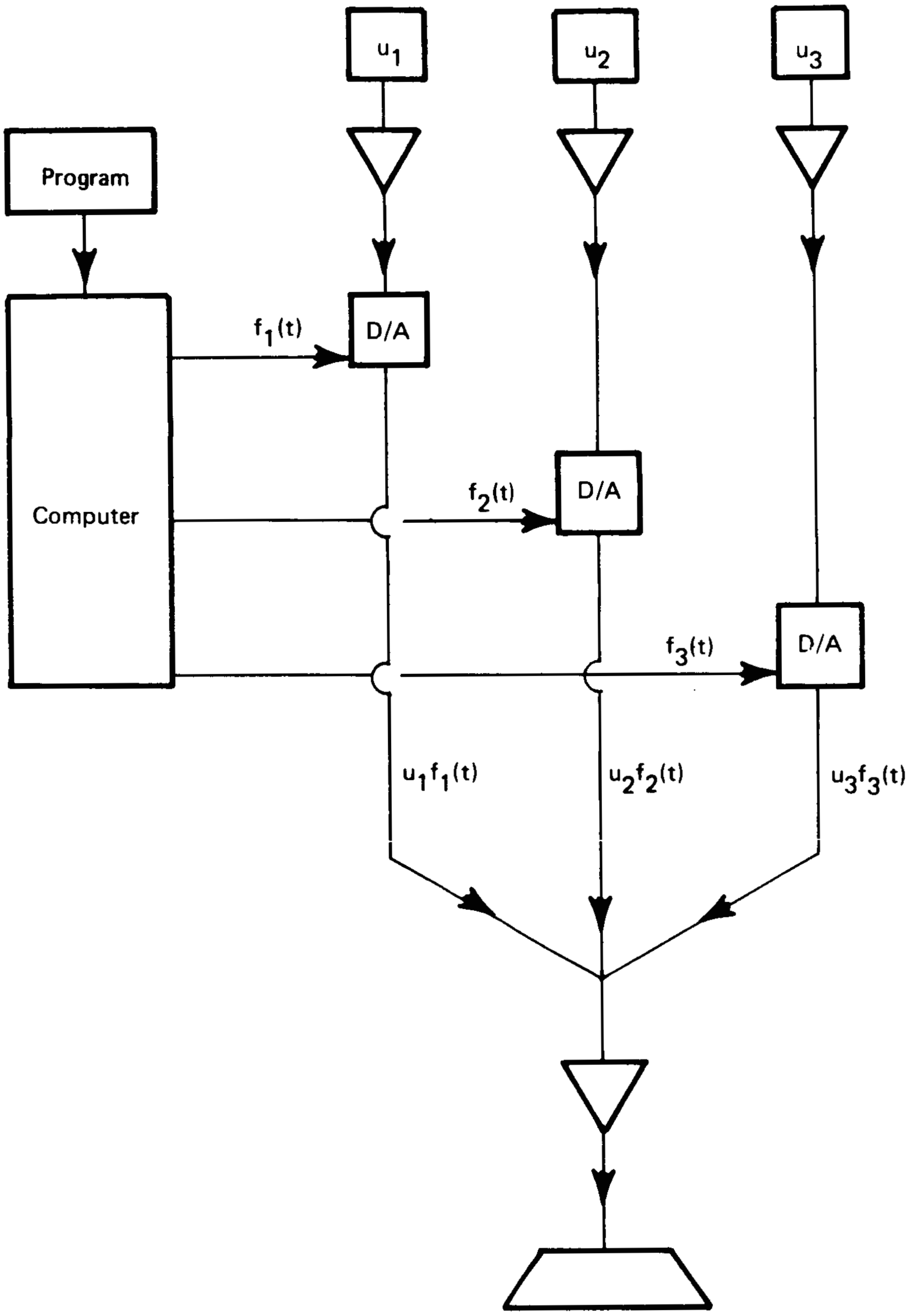


Figure 6. Time-dependent behavior of consonants *p* , *b* , *v* , and *w*. The solid line controls the spectral distribution belonging to *p* , whereas the broken line controls the vowel *a* . As the time-function flattens out, the perception shifts from *p* to *b* , then to *v* and *w*.



$$f(\nu,t) = u_1f_1(t) + u_2f_2(t) + u_3f_3(t)$$

FIGURE 7 Speech Synthesizer, including digital to analog converters

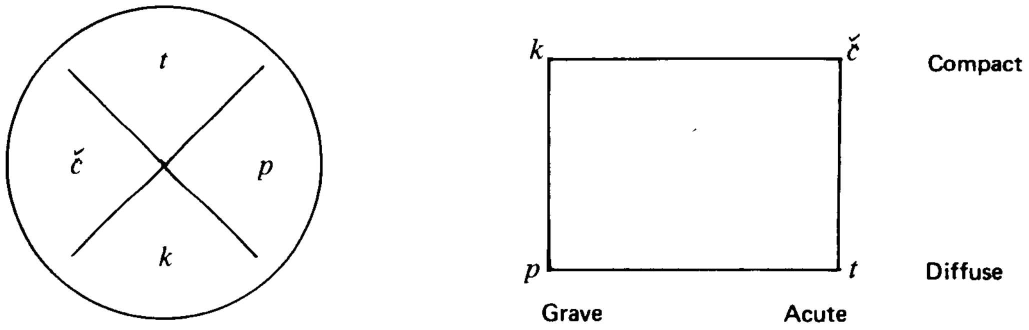


Figure 8, Distinctive-features approach tends to a binary separation of phonemes. It does not predict the transformation and contrast effects but may perhaps be justified within the perceptual theory.

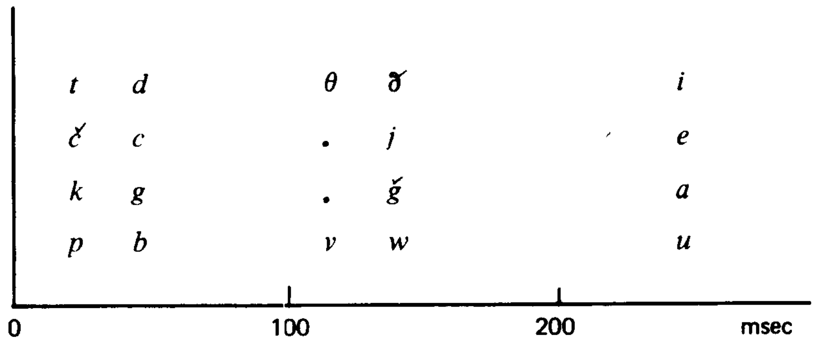


Figure 9. A two-dimensional array containing some consonants and their related vowels. Such organizations are representable (up to some transformations) in terms of a spectro-temporal function, $S(\nu, t)$.

LITERATURE

Benninghof, W., (Doctoral Dissertation, in Progress, Speech Communications Center, Northeastern University, Boston, Mass.).

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., and Gerstman, L.J., 1952. 'Some Experiments on the Perception of Synthetic Speech Sounds." J. Acoust. Soc. Am., 24, 597-606.

Cooper, F.S., Liberman, A.M., and Borst, J.M., 1951. "The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech." Proc. Nat. Acad. Sci., 37, 318-325.

Fant, G., 1967. "Auditory Patterns of Speech." Models for the Perception of Speech and Visual Form. W.Wathen-Dunn, ed., Cambridge & London: M.I.T. Press, 111-125.

Hiramatsu, K., Wackerbarth, R.K., and Coates, C.L., 1967, "Classification of Phonemes by the Distinctive Features--A Computational Approach." Conference Preprint, 1967, Conference on Speech Communication and Processing, M.I.T., (Office of Aerospace Research, U.S. Air Force.) 78-82.

Jakobson, R., Fant, C.G.M., Halle, M., May 1952, Technical Report No. 13, Acoustics Laboratory, M.I.T., (Cf. also authors' Preliminaries to Speech Analysis: The Distinctive Features and their Correlates. Cambridge: M.I.T. Press, 1965.

Jakobson, R., 1966. "The Role of Phonic Elements in Speech Perception." XVIIIth Intern. Congr. Psychology. Symposium 23: Models of Speech Perception. Moscow (Aug. 8, 1966).

Liberman, A.M., Cooper, F.S., Harris, K.S., and MacNeilage, P.F., 1963. "A Motor Theory of Speech Perception." J. Acoust. Soc. Am., 35, 1114. (Cf. also, with same title, in Proc. Speech Communication Seminar, Vol. 2. Stockholm: Royal Institute of Technology, 1962.)

Stevens, K.N., and Halle, M., 1967. "Remarks on Analysis by Synthesis and Distinctive Features." Models for the Perception of Speech and Visual Form. W.Wathen-Dunn, ed. Cambridge & London: M.I.T. Press, 88-102.

Winckel, F., 1967, Music, Sound and Sensation (Tr. by T.Binkley) N.Y.: Dover Publications, Inc., 121.

Yilmaz, H., 1962, "On Color Vision and a New Approach to General Perception." In E.E.Bernard and M.R.Kare, eds., Biological Prototypes and Synthetic Systems. Vol. 1, N.Y.: Plenum Press, 126-141.

Yilmaz, H., 1967a. "Perceptual Invariance and the Psychophysical Law." Perception and Psychophysics, 2, Nov., 1967.

Yilmaz, H., 1967b. "A Theory of Speech Perception." Bull. Math. Biophysics, 29, Dec., 1967.