Peter Ladefoged

working papers

in phonetics 4

university of california, los angeles

july 1966

UCLA

Working Papers in Phonetics 4

July 1966

The Phonetics Laboratory is part of the Linguistics Department, University of California, Los Angeles. The work reported here was also supported in part under USPHS Grant NB 04595 and in part through the UCLA Center for Research in Language and Linguistics.

Introduction

This is the second issue of Working Papers in Phonetics to appear in the current academic year. In general we hope to continue publishing in the early part of each summer papers reflecting work done in the previous year. These papers are designed to be partly a progress report on our sponsored research, partly a reasonably up-to-date account of our activities for the benefit of friends and colleagues working in similar areas (whose comments are always welcome), and partly a record which we use for our file and for the instruction books and service manuals used with our own insturmentation. We will also continue to issue doctoral theses relevant to our sponsored research (such as "Some phonetic specifications of linguistic units: an electromyographic investigation" by Victoria Fromkin, which came out as Working Papers in Phonetics 3) and other material on which prepublication comment is sought (such as the draft of a monograph on "Linguistic Phonetics" by Peter Ladefoged, which is in preparation). But we do not intend publication of material in this series to replace publication through more regular channels. We hope that there will never be any need for us or anybody else to give a reference in a published paper to anything published in here. Material worthy of regular publication will appear as soon as possible elsewhere. Revised versions of four of the main papers in previous issues have already appeared or are in press; and there are published abstracts of other papers on work which is still in progress.

Working Papers in Phonetics 2, which appeared at about this time last year, contained a brief survey of the functions and facilities of the UCLA Phonetics Laboratory. The functions remain substantially the same, although there have been some nominal changes. With effect from 1 July 66 the Laboratory became administratively part of the newly constituted Linguistics Department; but it will remain closely associated with the graduate programs in Speech and in English. The facilities have been considerably increased during the year. Some of the papers which follow describe the construction of a terminal analog speech synthesizer, the development of new photoelectric techniques for studying the state of the glottis, and a system for making recordings in the field of signals such as the pressure of the air in the mouth. Other instrumental developments include the setting up of the master and slave units of the Visible Speech Translator (supplied by Bell Telephone Laboratories) in two separate rooms so that it is possible to learn to communicate through the instantaneously produced spectrographic patterns (but none of us has as yet succeeded); the construction of a stroboscopic system for viewing the glottis; and the purchase of a second Sona-Graph and an Ampex stereo studio recorder. A grant from NSF has also allowed us to order an expanded LINC-8 general purpose computer, which we hope will be installed in the laboratory in October.

Relationship between Linguistic Units and Motor Commands

Victoria A. Fromkin

[Paper given at the 1966 Meeting of
the Acoustical Society of America]

The semicontinuous acoustic signal that we call speech is the result of a number of discrete neuromuscular events. We cannot store in our brain motor commands for every utterance that we may wish to say, owing to the brain's finite storage capacity. Speech then is produced by the rearrangement of a limited number of stored items. What is the size or nature of these stored units? How do speakers encode a sequence of discrete linguistic units into a continuously changing speech signal? An electromyographic study revealed that no simple correspondence exists between a phoneme and its motor commands. Initial and final allophones of consonants differed; both were unaffected by the adjacent vowels. Vowels were influenced by the preceding consonant. The results of this study relevant to the relationship between motor commands and phonetic segments are discussed. Alternative hypotheses relating to the nature of the stored unit are suggested.

Rules of Vowel Duration in American English

Chin W. Kim

[Paper given at the 1965 Christmas Meeting of the Linguistic Society of America]

In the introduction the paper discusses the scope of phonetic specification, i.e., how detailed and specific should the rules of the phonological component of a grammar be in converting the abstract representation of morphemes at the level of classificatory phonemics into real sounds at the level of physical phonetics. Pertinent statements by various linguists are reviewed and discussed. The writer's own view is then given.

This view is illustrated with rules of American English vowel duration. Experimental data show that the length of English vowels ranges from 100 msec to 400 msec, and that there are four factors that influence the length; (1) the tenseness of the vowel, (2) the degree of openness of the vowel, (3) the voicing of the following consonant, and (4) the manner of articulation of the following consonant. The paper discusses the degree and the nature of effects of these features on vowel duration, i.e., how much does each feature influence the vowel length; and which effects are contingent upon physiological constraints of the human vocal mechanism (and therefore not relevant in phonetic specification), and which are language-dependent, hence parts of the system of English (and therefore must be accounted for by the phonological rules of English grammar).

The rules are then formulated so as to assign appropriate values of length to vowels. Implications of these rules are discussed, and finally, some examples that are explained by these rules are given.

Mechanico-Acoustical Speech Synthesis

Ralph Vanderslice, Peter Ladefoged and James Anthony

[Paper to be given at the 1966 American Speech and Hearing Association Convention]

A replica of one subject's vocal tract above the glottis was constructed and tested with various horn drivers and airstream modulators which were adapted or constructed to simulate the glottal source. Steady-state vowels were produced, and techniques and mechanisms for providing dynamic tract-shape changes -- and perhaps ultimately connected speech -- were investigated. Measurements which are not readily accessible in the live subject were taken for each vowel tract-shape. The source spectrum (measured by removing the head) was compared with each vowel spectrum to yield the corresponding transfer function. The axial lengths and area functions were determined by taking a casting of each tract shape in elastic impression material, straightening it, immersing it by increments in a graduated cylinder, and measuring its fluid displacement to obtain the average area over each short segment. The resulting area functions were used as data for a computer program (Mathews and Walker, 1962) which calculated the transfer functions for each vowel. The computed functions were then compared with the measured ones -- partly to test the adequacy of the simplified theoretical model of the vocal tract embodied in the computations. The results suggest that for the detailed study of the relation between physiological-articulatory activities and acoustic output a complex model is needed. Parameters for dynamic control of tongue, jaw, velum and lip movements were hypothesized on the basis of data from sagittal X-rays and cineradiography and from electromyographic and photographic studies. Analytic expressions approximating the observed tract shapes were derived by digital computer using a polynomial regression program. Although it runs counter to the trend towards electronic analog or digital simulation of physical systems, mechanico-acoustical synthesis can add much to our understanding of the correlations between physiological and acoustic parameters of human speech.

An Attack on the Number Two

Peter Ladefoged

[Substance of a paper presented at a meeting of

the Acoustical Society of America, June 1966]

One of my main concerns, which I think is shared by many of you here in the Speech Communication section, is trying to give a precise account of the acoustic structure of speech. Ideally we would like to have a set of rules which related linguistic units with sounds, or at least with a defining set of acoustic properties. To a great extent, this has been done by Kelly at the Bell Telephone Laboratories, and by Holmes, Mattingley and Shearme who published the actual values required in a set of rules for synthesizing speech. But they were relating phonemic units with sounds. Now we must see whether we can state the rules in terms of acoustic features of phonemes, as was suggested a long time ago by members of the Haskins group in their preliminary work on speech synthesis by rule.

From a linguistic point of view, one of the best known sets of components of the phonemes is the set of Distinctive Features proposed by Jakobson, Fant, and Halle. I would like to consider the extent to which it is possible to define the acoustic characteristics of these features. Jakobson and his colleagues have suggested that the linguistic contrasts which occur within the languages of the world should be specified in terms of about 14 binary distinctive features, such as voice - voiceless, nasal - oral, etc. Their work is a great step forward, principally in that it formalizes the notion of phonetic specifiability and natural class of speech sounds, as presupposed by nearly all linguistic theories. But a number of difficulties remain. The Jakobsonian theory requires the features to be defined in terms of independent properties which have binary states but relative values. Everyone agrees that phonetic specifications must be in terms of relative values, since it is an obvious fact that what matters for communication within a language is not the absolute values of, say, the formant frequencies in vowels, but the relative values that occur in one vowel as opposed to another. The absolute values of acoustic characteristics such as the formant frequencies and the fundamental frequency may be significant for distinguishing among individual speakers, but they are not properties of the language.

I am not sure whether Jakobson, and Chomsky and Halle, consider their distinctive features to be completely independent. They at times speak of some features being primary in relation to others, which implies a hierarchy and not complete independence. Some of the definitions which have been given also require features to be considered relative to one another in that, for instance, the feature grave has been defined as having different physical correlates when applied to consonants as opposed to vowels. And there are other dependencies in that, for instance, the theory explicitly states that sounds cannot be both vocalic and strident. But their procedure for comparing phonological descriptions and evaluating one as being simpler than another does not involve any considerations of relationships of a hierarchical nature between features. They have never stated that the relationship between features has any theoretical status.

The binary nature of features has, however, been given considerable theoretical status. But in some cases this condition is impossible to maintain, if it is true that the properties of features are also relative and independent. Some linguistic contrasts, such as tones, consist of a number of items arranged along a single continuum. If it is true that the items are distinguished simply by the degree to which they have a given property, and if it is also true that the definitions of features must be expressed in relative terms, it is difficult to see how to deal with this situation by means of binary features. For example, let us consider a language (such as Yoruba) with three contrasting tones, high, mid, and low. We can distinguish between them by using two binary categories, high - non-high, and low - non-low. This is fine if the features are completely abstract, and are being used simply as labels for natural classes; any set of objects can be arbitrarily classified in this way. But it is not possible if the features have to be given properties. High can be defined as possessing a comparatively rapid rate of vibration of the vocal cords; and non-high as possessing a comparatively slow rate of vibration of the vocal cords. But low also has to be defined as possessing a comparatively slow rate of vibration of the vocal cords; and non-low as possessing a comparatively high rate of vibration of the vocal cords. Given the requirement that the features should be relative, so that high differs from non-high in a relative way and no arbitrary reference points may be used, and given that only one continuum is involved, so that high, mid and low differ only with respect to one property, it is logically impossible to define a distinction between these two features. In writing phonological rules we may want to group the tones by means of binary divisions [this point was very convincingly made with respect to Yoruba by Halle in the discussion after the paper]; and a ternary system may be very inconvenient. But if we are not playing games and indulging in hocus pocus linguistics, and if we want our descriptions in terms of classificatory features to be mappable onto phonetic data, then we cannot logically have two independent relative binary features within one continuum. It is of course possible to maintain that there are binary perceptual classes which are not related in any simple way to the physical data. We will return to this point later; but we can note here that there is no evidence that people the world over regard pitch distinctions among speech sounds as being ordered in terms of a number of binary categories.

The situation is even worse if we are trying to describe four objects distributed along one parameter, as in a tone language such as Tiv; and it is not improved by the use of Wang's suggested features high - low, and mid - non-mid. It is impossible to define Wang's terms, given the requirements that features should be specified by relative properties, and that features are not ranked relative to one another. Furthermore, no binary system reveals the fact that high is related to mid-high in the same way as mid-high is related to mid-low, and in the same way as mid-low is related to low. This kind of relationship is important in many tone lowering rules. And presumably we want a theory of phonetic description and an associated evaluation criterion that enables us to say that a rule lowering each of three tones by one degree is more general than a rule which says that some of these tones change one way and others another. But in a binary system this is not so.

The same arguments can be used to show that there are weaknesses in specifying vowels in binary terms. At the moment the terms in use

able

t

ures

ed,

ures.

ıe

are compact, which means a high first formant, as opposed to non-compact, which means a lower first formant; and diffuse, which means a low first formant, as opposed to non-diffuse, which means a higher first formant.

If the definitions are related so that the definitions of compact and non-diffuse overlap (or, putting it another way, if the definitions preclude the possibility of a sound being both compact and diffuse; and it is clear that this is so) then we do not really have a binary system. And there is no reason why we should. It is perfectly possible to have a ternary (or n-ary) system in which items are classified as having one of a number of ranks of an identifiable property, and a simplicity metric which defines natural classes preferentially by rank and then by feature differences.

Put simply, the argument I am discussing here is that if sounds differ in and only in a single property on a given line, then it is impossible to divide them into more than two independent but relative groups. You cannot make a useful second division on the line without knowing where the first one came.

There are signs that the proponents of Distinctive Feature theory are realizing some of these difficulties, and are claiming that the features cannot be defined in terms of physical (acoustic or physiological) properties, but only in terms of perceptual properties. It is difficult to know what this might mean. If it is supposed to be a testable statement, I cannot see what test I can use to find out how untrained people group speech sounds which is not dependent on the linguistic background of the subjects; and if we use trained phoneticians as subjects, so that we partially remove the influence of the subject's linguistic background, then we might just as well remove all such influences and devise classes based on the physical and not the perceptual properties of speech sounds. If a particular perceptual feature has any universal validity it must be due to some property of the physical stimulus.

Chomsky and Halle have been foremost among those pointing out that linguistic descriptions must involve a phonetic component which is independent of the language under description. Perceptual groupings which simply reflect the linguistic competence of the subjects can never be used as the basis for a universal phonetic alphabet without destroying the whole concept of natural class and phonetic specifiability. I would in fact go further and suggest that phonetic specifiability, when applied to phonological descriptions, should mean that it is possible to generate testable physical events. We must have a theory which allows us to map descriptions onto observable data. And the Jakobsonian theory of binary but relative features does not allow us to do this. Even though the phonetic matrix proposed by Chomsky and Halle may now include multi-valued entries, Jakobsonian distinctive features cannot be used to categorize sounds in the underlying classificatory matrix because the matrix they define cannot always be given a phonetic interpretation. But, as Chomsky and Miller have said: "It is an extremely important and by no means obvious fact that the distinctive features of the classificatory phonemic matrix define categories that correspond closely to those determined by the rows of the phonetic matrices." Insofar as Jakobsonian features cannot be related to phonetic categories, the theory is invalid.

Terminal Analog Speech Synthesizer

Norris P. McKinney, Marcel A.A. Tatham and Peter Ladefoged

A speech synthesizer has been constructed as part of a project involving the development of different synthesis systems. In the last two years we have gained a great deal of experience with static models of the vocal tract. In the future we hope we will be able to construct dynamic analogs operating in terms of articulatory parameters. Meanwhile we wanted to gain some experience in operating dynamic controls and circuits; and we needed a facility for synthesizing speech in connection with various research projects, including studies of the perception of speech-like stimuli, and of the formulation of rules connecting the output of the syntactic component of a grammar with observable phonetic data. We therefore decided to construct a simple speech synthesizer with the ability to generate, fairly accurately, acoustic signals corresponding to speech. The synthesizer is of the 'terminal analog' type; i.e. in producing an output similar to the vocal tract output, only one isolated resonant circuit is used in cascade for each important vocal tract system resonance. This is in distinction from the 'line analog' type of synthesizer, in which the synthesizer system resonances are obtained by means of a relatively large number of coupled resonant circuits in cascade, one for each incremental section of the vocal tract. Control signals are derived from lines drawn in conducting ink on a moving plastic belt. In the design and construction of the system considerable benefit was derived from the experience of researchers at the Communication Sciences Laboratory, University of Michigan and in the Phonetics Department of the University of Edinburgh.

General Description

The system consists of a pulse train with a variable fundamental frequency, three formant resonators with individually variable resonant frequencies and amplitudes, a fourth formant resonator of fixed frequency and with an amplitude controlled by that of formant three, a noise generator with variable amplitude and variable center frequency. Nine control voltages are used to vary: (1) the repetition rate of the pulse corresponding to that produced at the glottis in the human vocal mechanism, and having a spectrum falling at a rate of 6dB/octave above 100 Hz; (2), (3), (4) the amplitudes of the formants; (5), (6), (7) the center frequencies of the formants; (8) the amplitude of the noise generator; and (9) the center frequency of the noise band.

Larynx Pulse Generator

The larynx pulse generator (Figure 2) uses a voltage to frequency converter. The pulse output frequency of this circuit is proportional to the control voltage (0 - 10v) over the range 55 - 275 Hz. This pulse is used to trigger the monostable multivibrator of the larynx pulse generator, producing a 100 microsecond pulse at its output. This signal has a spectrum envelope which drops to 3dB below its low-

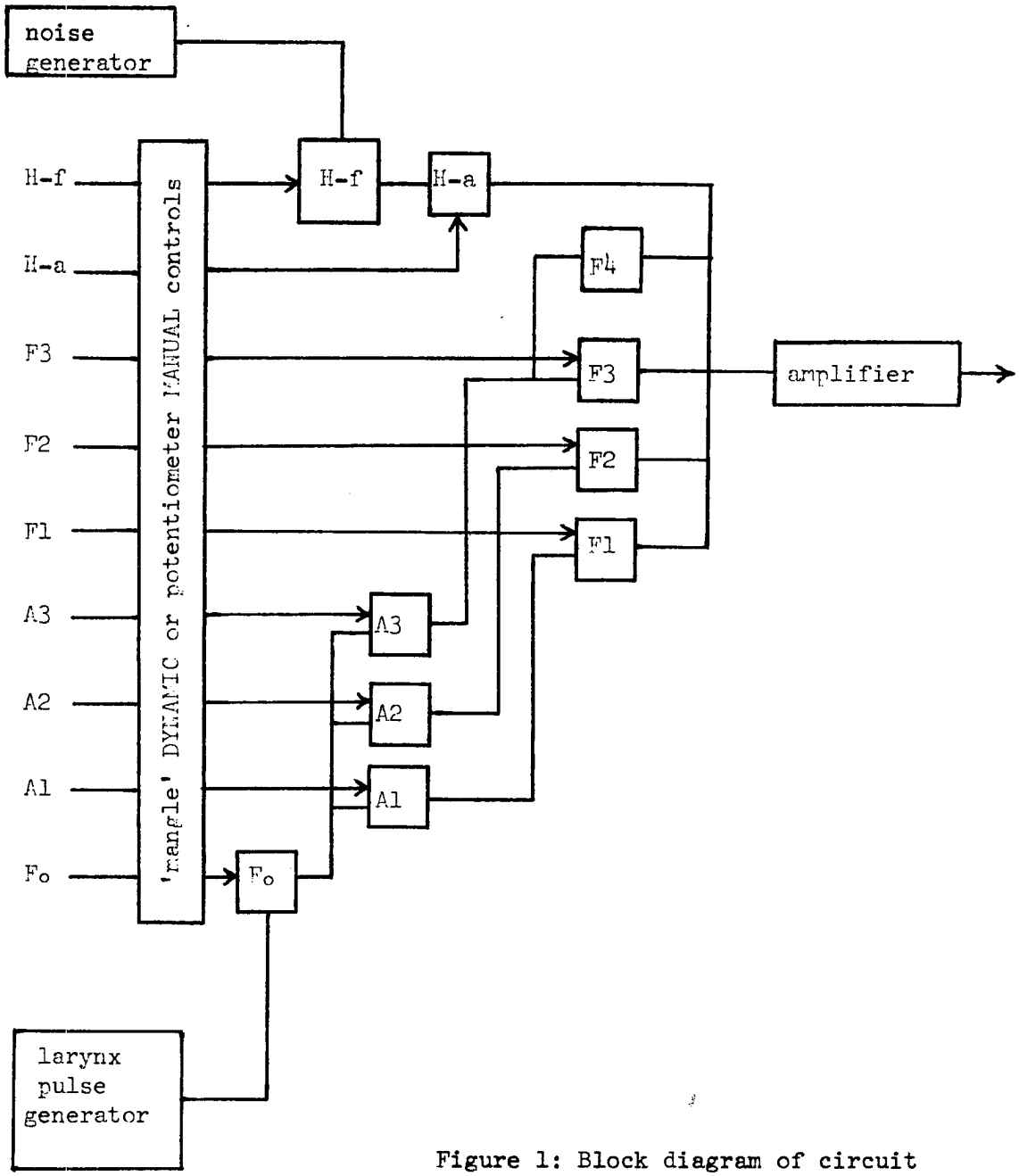


Figure 1: Block diagram of circuit

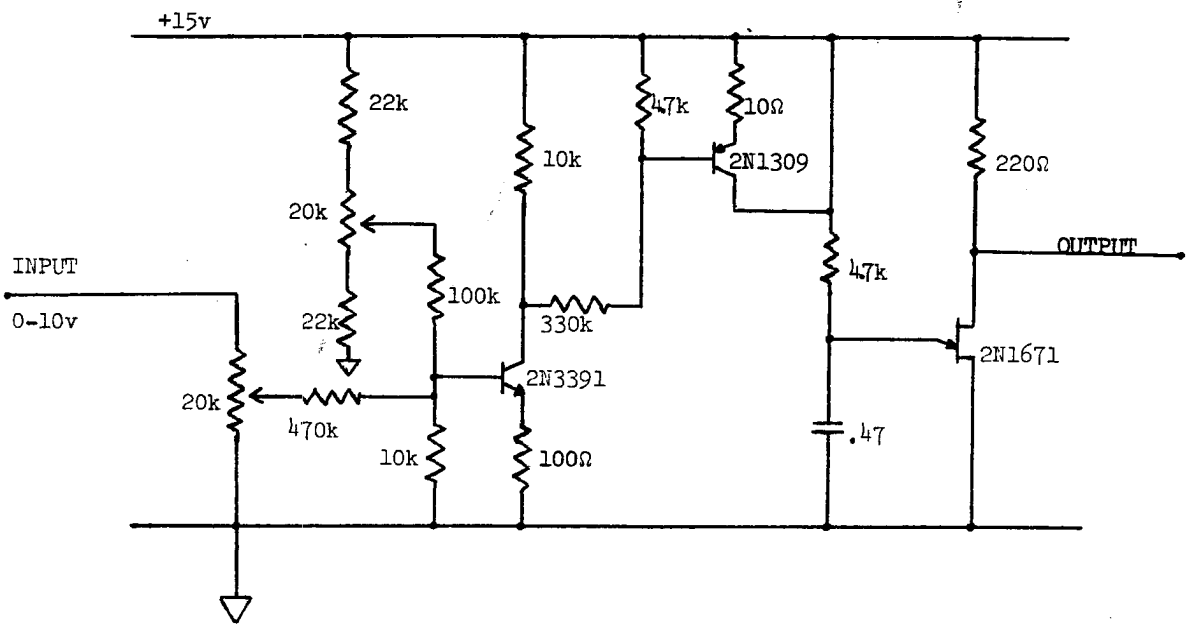


Figure 2 (a): Voltage to frequency converter

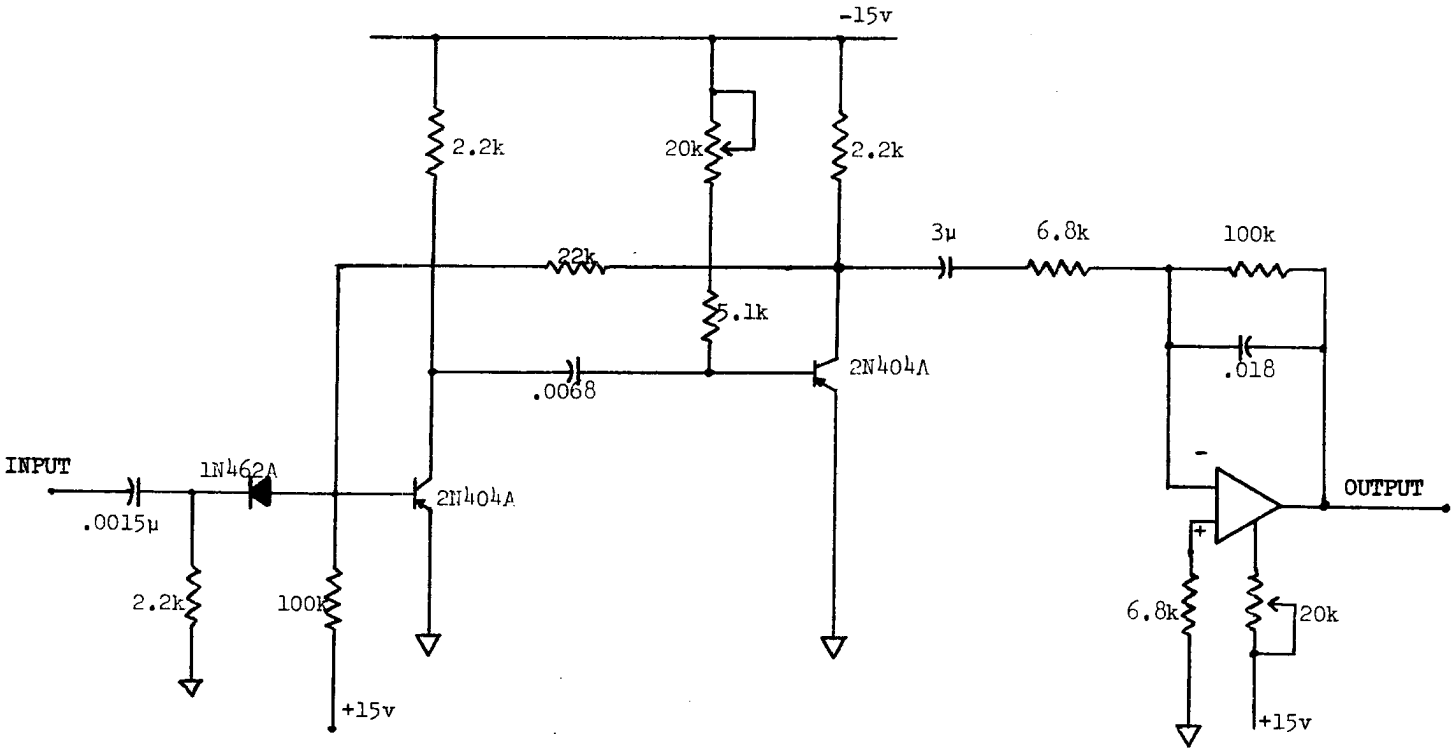


Figure 2 (b): Larynx pulse generator

frequency asymptote at 4.5 kHz. The signal is then filtered and amplified (with dc blocking) to form a pulse train whose spectrum falls at 6dB/octave above 100 Hz.

Gain Control Circuits

The output from the larynx pulse generator serves as input to three gain control circuits (Figure 3). Here the amplitude of the larynx pulse to be fed to the formant resonators is varied according to the control voltage applied. The circuits have been arranged to provide an average control characteristic slope of 5dB/v; however, the curve has been made to fall sharply when the control voltages go below 2v, in order to suppress the output signal below the output noise level wherever silence is intended. An additional gain control circuit is provided for the control of the signal level from the noise generator.

Formant Resonators

Despite the well known advantage of varying the capacitance in a speech synthesizer formant circuit, rather than the inductance, the availability of good quality saturable reactors (Vari-L inductors) made the latter alternative attractive. Model EL-215 Vari-L inductors were chosen for all the formant circuits. Only a few models of Vari-L inductors are designed for use in circuits resonating at the lower speech formant frequencies. Among these models, the EL-215 has superiorly low hysteresis and temperature coefficient of inductance change. The principal problem encountered in the use of the variable inductors in the formant circuits was their considerable tendency to be microphonic.

It is widely believed that the bandwidth of speech formants is not a factor of crucial importance in speech synthesis. For this reason, the bandwidth of the formants has been made to vary as a fixed function of the formant frequency. That is, no provision has been made for either dynamic control of formant bandwidth as one of the synthesis parameters, or for presetting the formant bandwidth. Instead, the formant bandwidths are made to vary with formant frequency according to relations between the two which have been published in the technical literature. The first formant bandwidth remains constant at 50 Hz. This relation has been suggested by Fant (1960), and corresponds closely to data published by Dunn (1961). For the relation between the resonant frequency and bandwidth of the second formant circuit the straight line approximation given in a figure in Dunn's paper was chosen. Seemingly inappropriate values of formant bandwidth resulted from extrapolating Dunn's straight line approximation of third formant bandwidth vs. frequency to cover the resonant frequency range desired for the third formant circuit. A modified version of this relation is realized in the synthesizer circuit.

JTPUT

The resonant frequency of the circuit formed by adding a fixed capacitor to the variable inductor is a near linear function of the inductor control current. A diode function generator inserted between each formant frequency parameter control voltage input and the corresponding inductor control coil compensates for slight deviations from linearity, resulting in an essentially linear relation between an input control voltage and the resulting resonant frequency of the formant circuit. A sufficiently

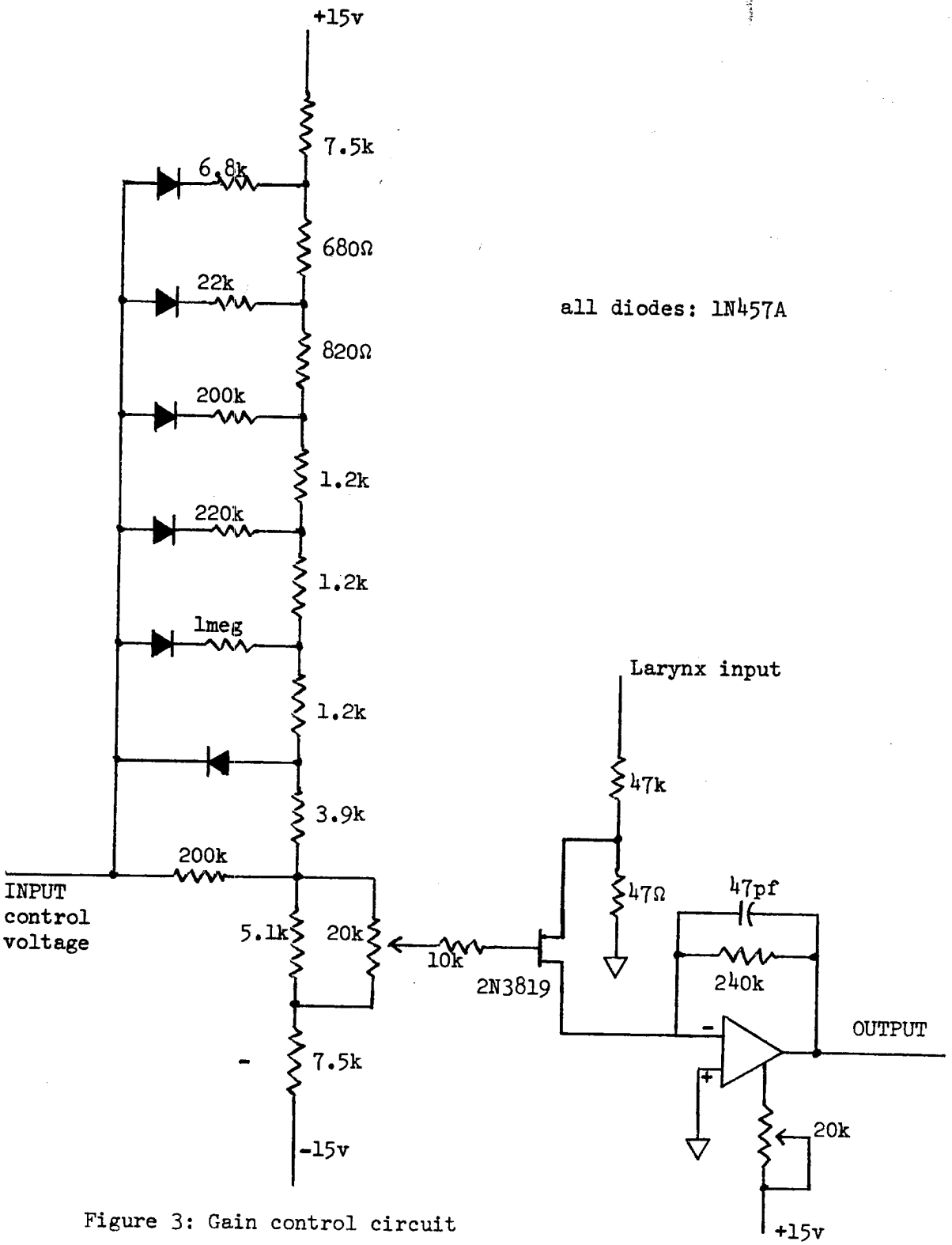


Figure 3: Gain control circuit

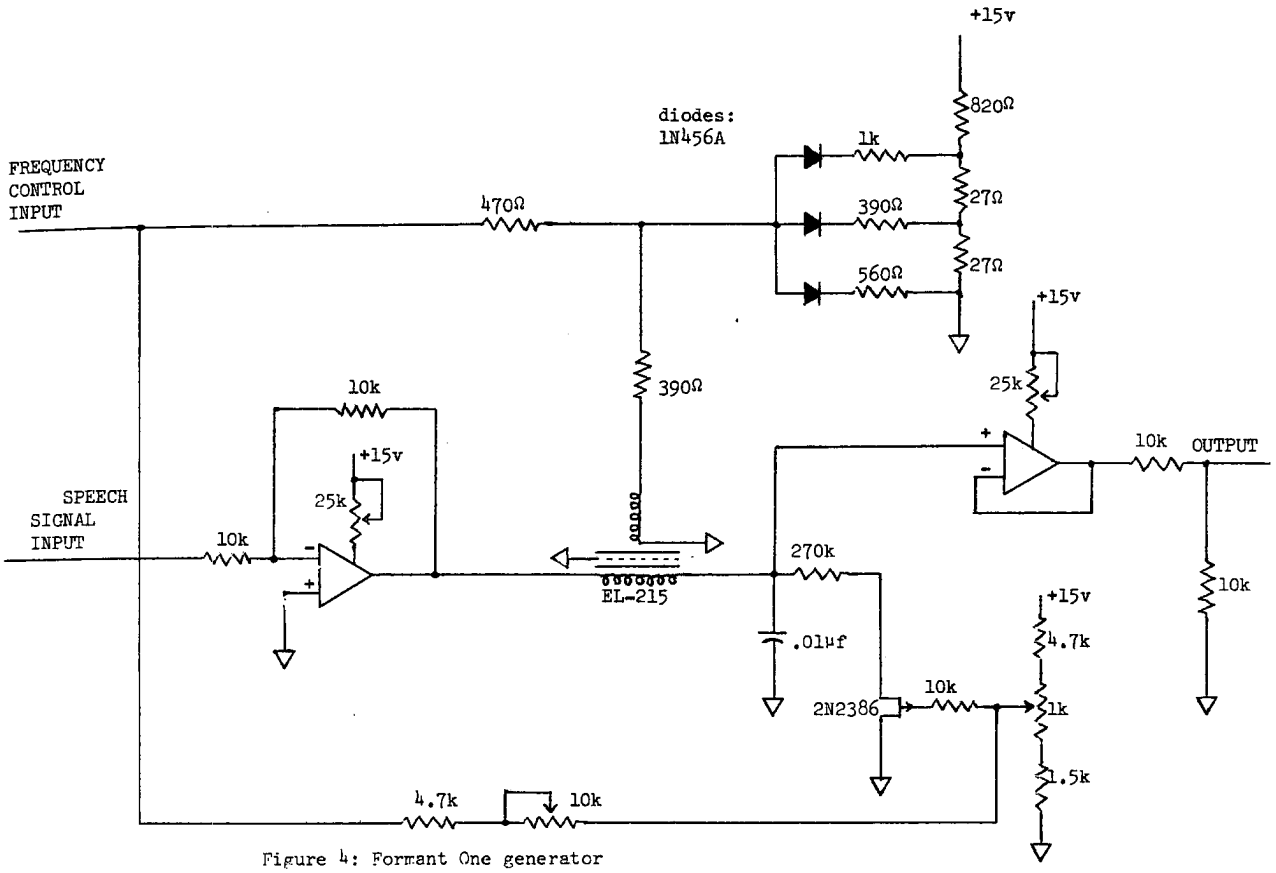


Figure 4: Formant One generator

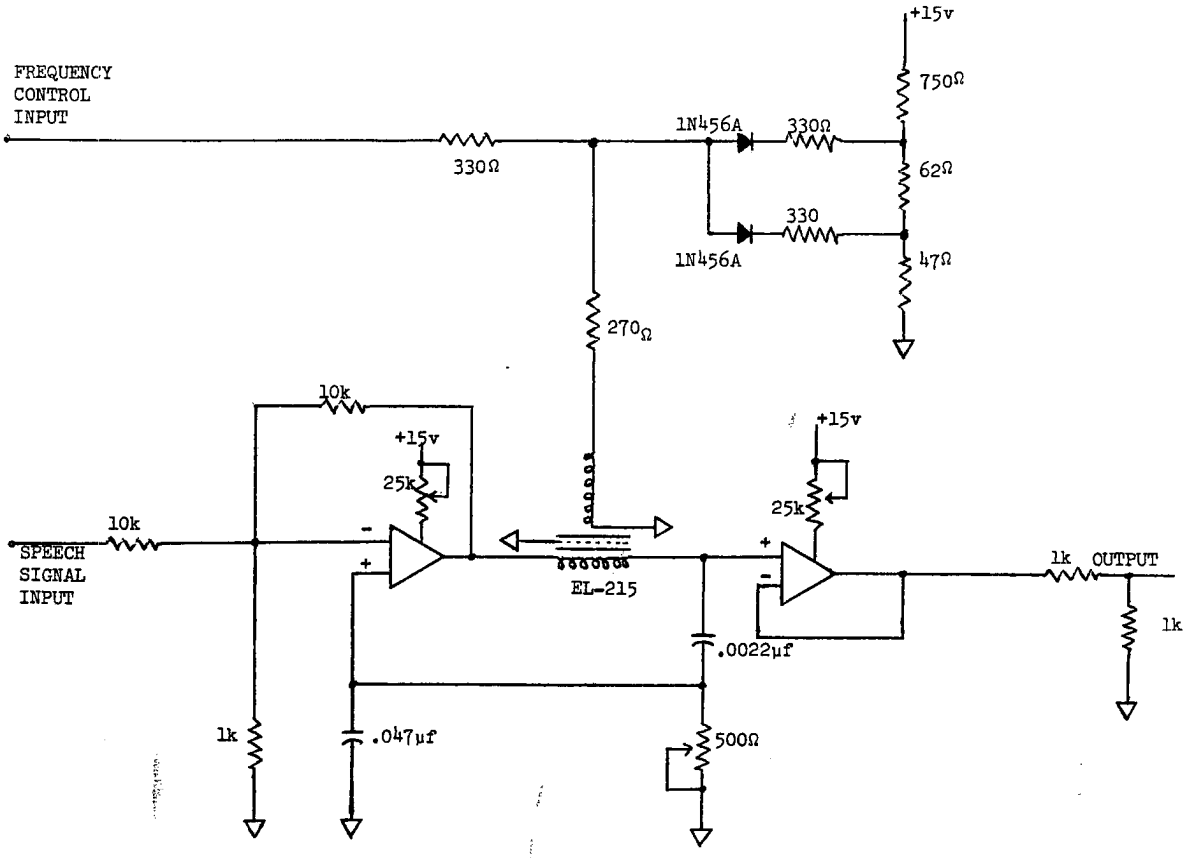


Figure 5: Formant Two generator

high Q to meet the design requirement for the first formant is obtainable with the basic circuit. After the addition of isolation amplifiers and a diode function generator as discussed above, only one additional circuit was needed to complete the first formant resonator. A variable resistance has been added across the capacitive element of the resonant circuit to reduce the Q (increase the bandwidth) to the specified value at each resonant frequency. A field effect transistor is used as the variable resistance, and its gate voltage is a linear function of the formant frequency control voltage.

The Q of the basic circuit could not be made sufficiently high to meet the requirement set by natural speech data for formants higher than the first. A feedback configuration, essentially similar to one used by Klatt at the University of Michigan, has been utilized to reduce the bandwidth of the transfer function for all formant circuits but the first. The feedback signal is obtained as a voltage across a small resistor in series with the capacitor; thus the feedback voltage is proportional to and in phase with the current in the basic series resonant circuit. A capacitor is used to reduce the feedback at higher frequencies.

In the third formant circuit it was necessary to employ a combination of the positive feedback and variable shunt resistance techniques in order to obtain the desired bandwidth as a function of resonant frequency.

Output

The outputs of the four formant resonators are combined in a summing amplifier (Figure 8) the output from which is fed through a power amplifier to the loudspeaker. The power amplifier is a commercially available high-quality unit (Fairchild Model 688).

Control

A series of switches on the front panel enable any or all of the parameters to be switched between dynamic and manual operation. Thus the synthesizer may be entirely in dynamic or manual mode, or in dynamic with manual control over any chosen parameter(s).

In the manual mode, the parameter control voltage is proportional to the manually set position of the corresponding potentiometer on the front panel of the synthesizer.

The purpose of the dynamic control unit is to provide continuously varying voltage inputs to the control circuits of the various parameters. The system employed is similar to the 'mangle' system used in the Edinburgh speech synthesizer, PAT (Anthony and Lawrence 1962). A belt of plastic is made to pass under a resistive roller across which 10v is applied. Thus lines drawn in silver conducting ink on the belt take off a voltage from the roller corresponding to their position on the belt. These voltages are then conducted through the belt, at the point where it is joined, to nine parallel lines drawn on the reverse side and picked up by nine brass strips fixed lengthwise with respect to the belt's direction of travel. The voltages thus derived from the resistive

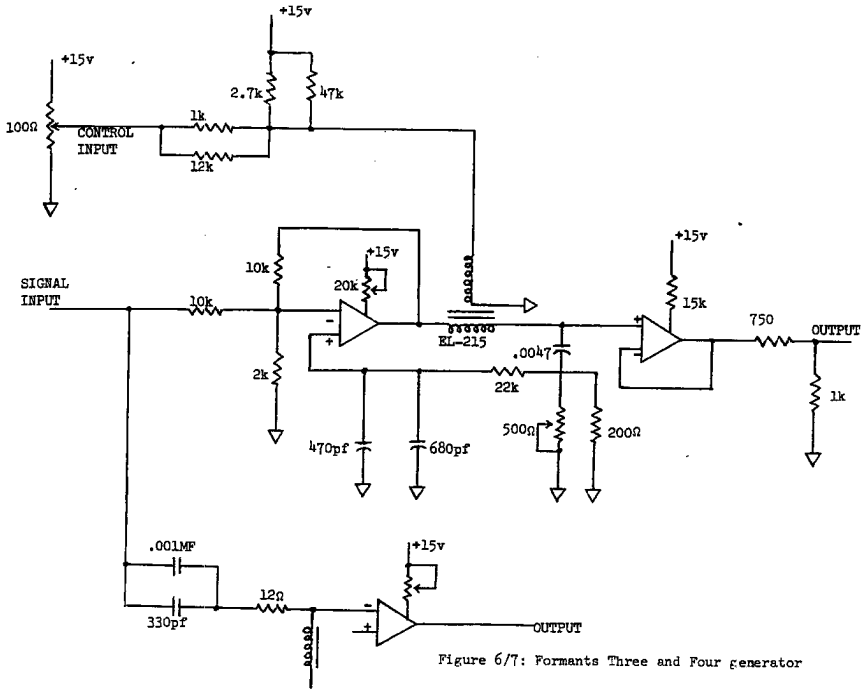


Figure 6/7: Formants Three and Four generator

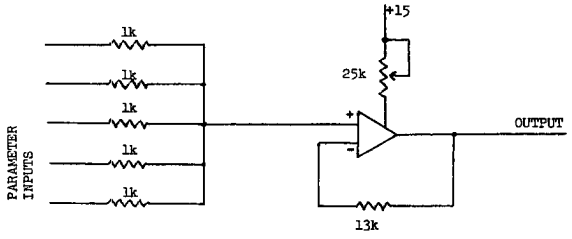


Figure 8: Summing amplifier

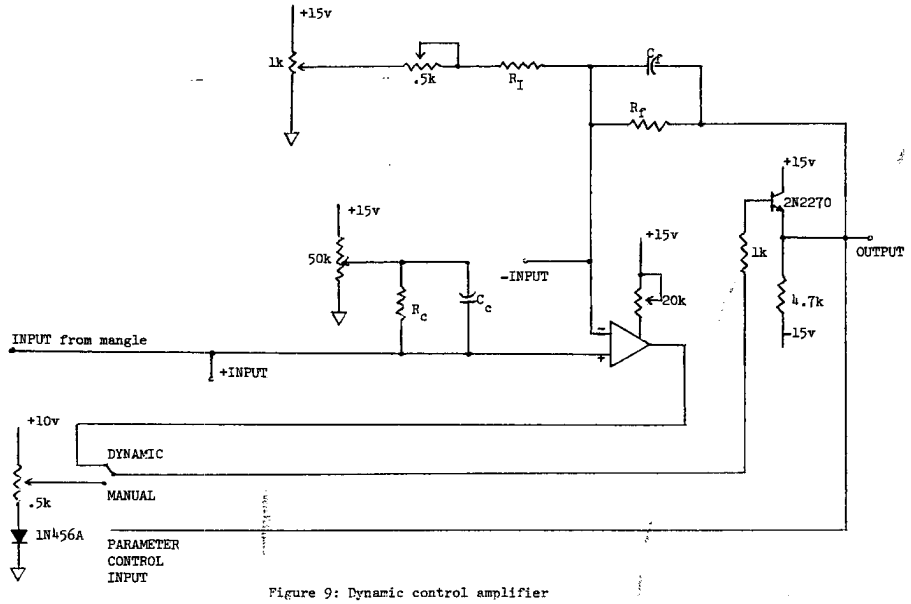


Figure 9: Dynamic control amplifier

roller are applied to the 'mangle' amplifier circuits.

The purpose of the linear isolation amplifier circuits (Figure 9) is to amplify and translate the 'mangle' output voltages and keep the 'mangle' output current at a very low level. Voltage amplification and translation of the 'mangle' output signals is necessary because the range of the parameter control input voltages (e.g. to the formant resonator circuits) has been standardized at 0 - 10v, while the range of each 'mangle' output signal (e.g. 3.2 - 4.1v or 8.3 - 9.8v, etc.), varies according to the zone of contact which the corresponding line has with the resistive rollers. Very high current amplification is required for isolation of the 'mangle'output lines because their connection to the resistive roller potentiometer through a rolling contact and a sliding contact has relatively high and variable resistance.

Construction

Construction is entirely in modular solid-state form using Vector 812 WE circuit boards held in two Elco Varipak II card cages; The entire unit has been mounted in a single 19" rack with the dynamic control 'mangle' attached. Extensive use has been made of commercially available solid-state operational amplifiers; this, together with the modular construction, makes for a certain measure of standardization, ease of construction and location of faults.

Acknowledgement

Our thanks are due to Mr. Willie Martin who wired most of the circuits.

Bibliography

Anthony, J., and Lawrence, W. (1962) 'A resonance analogue speech synthesiser' Proc. 4th. Internat. Congr. Acoustics.

Dunn, H.K. (1961) 'Methods of measuring vowel formant bandwidths' J. Acoust. Soc. Amer. 33, 1737.

Fant, C.G.M. (1960) Acoustic Theory of Speech Production 's-Gravenhage: Mouton.

Some Requirements for a Model of Performance

Victoria A. Fromkin

There has been much discussion in linguistic literature in recent times concerning the requirements and properties of models of linguistic competence. Competence is related to performance as 'langue' is to 'parole'. 'Langue' or 'competence' thus refers to the "underlying system of rules that has been mastered by the speaker-hearer" (Chomsky 1965), and 'parole' or 'performance' refers to the way a speaker-hearer utilizes this 'internalized grammar' when he actually produces utterances.

It is, perhaps, unfortunate for the development of an adequate theory of competence as well as performance "that the only studies of performance, outside of phonetics, are those carries out as a by-product of work in generative grammar" (Chomsky 1965). Even in phonetics there has been more concern with taxonomic descriptions of speech sounds, and too little concern with performance models. Until an adequate theory of linguistic performance is suggested, the relationship between competence and performance will remain vague. Furthermore, the differences between competence and performance will be confused, leading to inadequate theories of both.

A model of performance is a way of explaining the regularities in a speaker's behavior when he produces an utterance. Performance differs from competence in that the output of a performance model is determined partly by the rules of performance, and partly by the interference of external noise. But many of the properties of competence models apply equally to performance models. According to Chomsky (1965)

> every speaker of a language has mastered and internalized a generative grammar that expresses his knowledge of his language. This is not to say that he is aware of the rules of the grammar....a generative grammar attempts to specify what the speaker actually knows, not what he may report about his knowledge.

Just as a competence model need not specify what the speaker consciously is aware of, the performance model is not required to specify a speaker's conscious knowledge of how he produces utterances. But it must account for how a person utilizing his competence, his 'internalized grammar', actually speaks and the mechanisms that determine this.

It is also a requirement of a performance model to explain the "creative" aspect of speech. The rules which describe and predict performance must account for the ability of speakers to produce and understand utterances never spoken.

In light of the above, Katz and Postal's attempt (1964) to point out the differences between competence and performance fails to clarify in any meaningful sense the properties of either model. They compare a linguistic description of linguistic structure (a model of competence) to an axiomatic mathematical system. "In both cases the rules of the system simply define the notion 'derivation within the system'. The rules of a linguistic description no more describe how the speaker produces or understands sentences than the rules of a mathematical system describe the way in which proofs are written out or checked." (166)

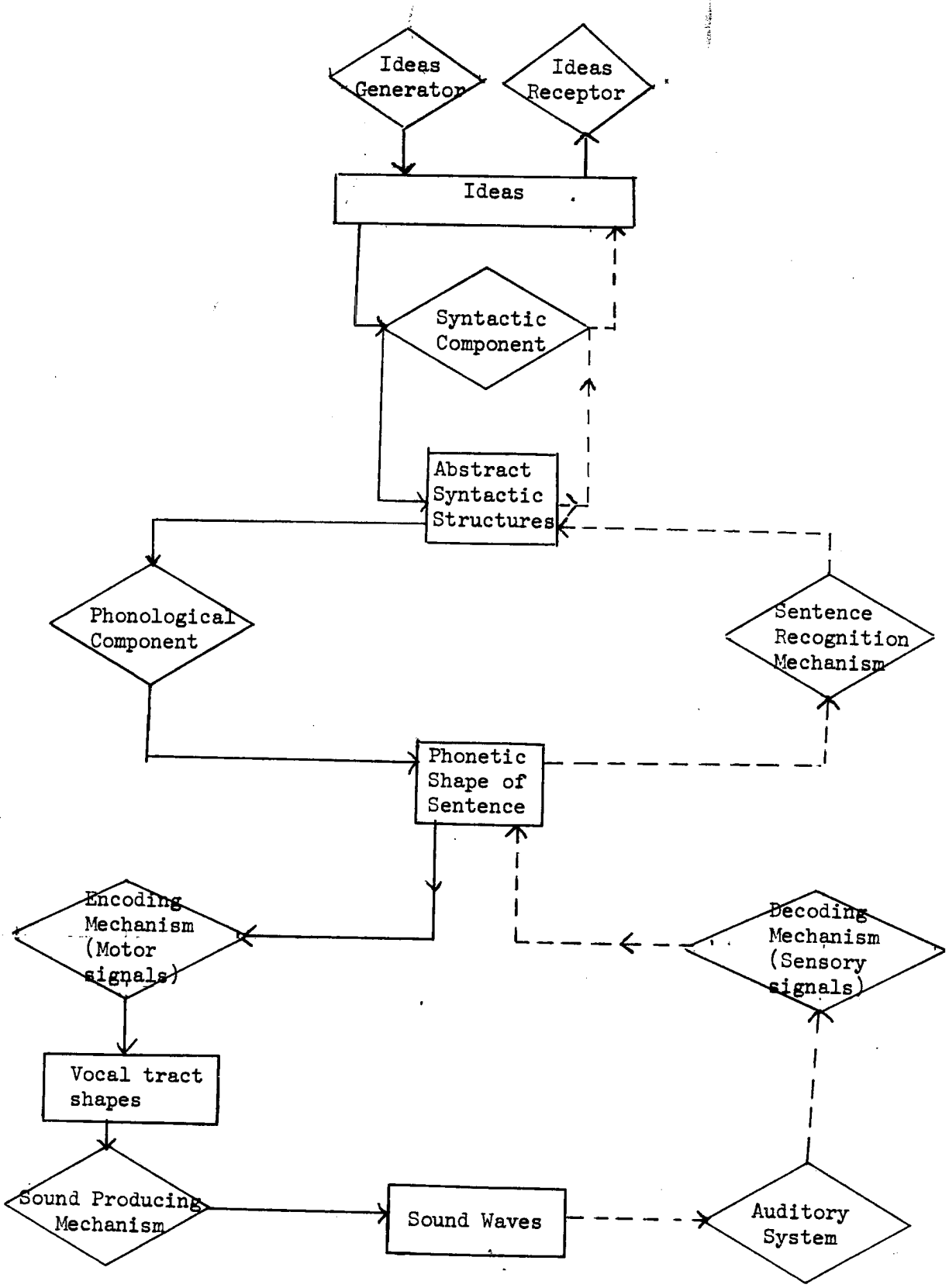


Figure 1 Schematic Diagram of Katz' Model (Katz 1964)

But a model of speech production (performance) can be just as formal a system as a model of competence, with its own axioms and rules, and its own theorems derived from them. Since both theories are physical theories in that they attempt to model real processes, both models must be isomorphic with the world of observable phenomena. The phenomena which they aim to describe and explain are different, thus necessitating different sets of axioms and rules.

Katz (1964) proposes a model which appears to be a model of both competence and performance.

> Given that both speaker and hearer are equipped with a linguistic description and procedures for sentence production and recognition, (the speaker starts with ideas he wishes to express)...uses the sentence production procedure to obtain an abstract syntactic structure having the proper conceptualization of his thought... After he has a suitable syntactic structure, the speaker utilizes the phonological component of his linguistic description to produce a phonetic shape for it. This phonetic shape is encoded into a signal that causes the speaker's articulatory system to vocalize an utterance of the sentence. (132)

The model also includes the decoding aspect on the part of the hearer. This suggested model can be diagrammed as in Figure 1.

Katz actually posits that the boxes correspond to separate components in the brain.

> Within the framework of the above model of linguistic communication, every aspect of the mentalistic theory involves psychological reality. The linguistic description and the procedures of sentence production and recognition must correspond to independent mechanisms in the brain. Componential distinctions between the syntactic phonological and semantic components must rest on relevant differences between three neural submechanisms of the mechanism which stores the linguistic description. The rules of each component must have their psychological reality in the input-output operations of the computing machinery of this mechanism. The ordering of rules within a component must...have its psychological reality in those features of this computing machinery which group such input-output operations and make the performance of operations in one group a precondition for those in another to be performed. (133)

Thus Katz' model is a model of performance, since this is what he says must happen when we communicate. But as such it fails.

When people speak they do not speak in 'grammatical' sentences, or sentences which they know to be grammatical. Actual speech is replete with false starts, grammatical deviations, slips of the tongue, etc. A performance model for speaker and hearer would have to show that, for one thing, only parts of sentences are encoded into motor commands

before the entire sentence is produced. How else would a speaker say "The boy --uh--I mean the girl--oh you know who I mean--went to Bost--no, New York."? But neither can the Katz model be a competence model. The proponents of competence models do not suggest that in speaking a speaker goes through all the rules which he has internalized to produce a sentence. "Generate" does not mean produce. (Chomsky, 1965)

The above model is inadequate either as a model of performance or competence; the construction of an adequate **performance** model should help to explain the process of speech production. Such a model must meet the general requirements of any scientific model, and also the specific requirements dictated by the data under consideration.

General Requirements

It is necessary to first clarify our use of the word 'model', and the requirements for any such model. One notion of model is used in pure mathematics, and refers to an interpretation of a formal axiomatic system. The starting point is not a set of physical concepts, or a physical process, but a formal system consisting of a set of primitive elements, a set of axioms, and a set of derived theorems. A model of such an axiomatic system is some realization or interpretation of it. There may exist many isomorphic models for any one formal system.

This is not the concept of model which will be discussed here. Rather, we shall employ the term 'model' to mean some representation of a physical process. We therefore start at the point where the mathematician may end. Such a model may be a physical replica, such as a ship model. It may also be a working apparatus which behaves similarly to the original system being modeled. Or it may be a theoretical model of a physical process such as Bohr's model of the atom which attempts to mirror the process as it is theoretically conceived. A mathematical model of a physical process may be constructed which attempts to capture some of the properties of the original physical system in precise and explicit terms. Newton constructed such a mathematical model of the planetary motions, obtaining thereby his law of gravitation.

Our concern in this paper is to suggest some of the requirements of a simple model of performance attempting to account for the motor commands which constitute the articulatory program for an utterance of speech. Such a model would be a theory, a set of hypotheses. From these hypotheses we wish to be able to predict events and to determine by experiment whether observation confirms these predictions and thus the initial assumptions. The purpose, then, of any such model is to explain the phenomena and it is justified in so far as it does make events understandable.

Unlike the **mathematician's model**, we start from a collection of physical data and end with the physical data. The physical data, however, are important only in so far as they are related to an explanatory theory.

> Surveys, taxonomy, design of equipment, systematic measurement and tables, theoretical computations--all have their proper and honored place, provided they are parts of a chain of precise induction of how nature works. All too often they become ends in themselves. (Platt, 1964)

In the construction of models of complex physical phenomena, it is often advisable to single out certain aspects of the phenomena, rather than attempt to model the totality. To attempt even this, we must gather as much data as possible. The validity of any theoretical model stems from an accurate investigation and coordination of a large number of facts. Where facts are so vague and poorly established as to be refractory to the scientific method, we generally find that there are so many different ways of coordinating them loosly that almost any opinion can be expressed with the same degree of plausibility. Yet, it must be remembered that an experiment repeated fifty times cannot give us any positive assurance that the result will be the same when the experiment is performed once again. If we are ultra cautious and refuse to generalize, experiment becomes useless since it cannot be used as a source of more general knowledge. Even in the simplest experiment, assumptions cannot be avoided; ordinary generalizations, without which science could not proceed, rests on an assumption.

An initial assumption is valid only if it is capable of being refuted. As W. A. H. Rushton has stated "A theory which cannot be mortally endangered cannot be alive." (Platt 1964) Since one can never prove an hypothesis, the responsibility of the scientist is to subject each hypothesis to rigorous tests. "It must be possible for an empirical scientific system to be refuted by experience." (Popper, K.R. 1959)

This approach is generally accepted in the scientific community. Yet, in linguistic literature one finds a seemingly opposite viewpoint. Saporta (1965) has stated that

> if one adopts the view that such notions [phonemes, nouns, etc.] constitute terms in an overall theory of considerable abstraction, it is difficult to know how one would go about asking for behavioral **correlates** for a given part of such a theory, in any way that could be said to constitute a crucial test .... If a test designed to demonstrate behavioral correlates for ... [such notions] fails to yield the predicted results, one feels obliged to modify the test, not the theory. (98)

Obviously, tests can be devised which do not actually test the validity of the hypothesis under consideration. One should hardly have to mention such procedures since they are not tests at all. Nor are operational and behaviorist tests required. If Saporta's attack is against strict behaviorism, or the philosophy of operationalism he should make this clear. Due to his lack of clarity, one could infer from his statement that there are some hypotheses which cannot be tested, but somehow are "true" in and of themselves, for all time. Such hypotheses belong to the field of metaphysics, not to science. This is not to deny that at times in the history of science the concepts suggested by a new theory may be so startling and so advanced in relation to the current technology that procedures for testing may not be forthcoming for a lengthy period of time. Perhaps the classical separation of theory and applicable experimental tests occurs in the work of James Clerk Maxwell in his theory of electro-magnetism, wherein some fifty years elapsed before Heinrich Hertz could demonstrate the existence of the phenomena contained in Maxwell's theory. Maxwell's theory was not incorrect because it suggested no immediate procedures for testing it. But in this sense Maxwell's hypothesis was extremely speculative

and his theory received serious attention and acceptance, only when Hertz, by direct experiment, corroborated one of its major anticipations (electromagnetic waves). This is the important question which Saporta seems to have missed. If the predictions suggested by the theory are negated by experimental tests, the theory indeed must be revised or discarded.

On the other hand, a model which predicts everything cannot be be tested, because it cannot be refuted. Actually, such a theory predicts nothing. If, for example, it is possible to predict from a model of linguistic performance that people produce sounds in terms of phonemes, or distinctive features, or syllables, or words etc, it would be trivially correct and would tell us nothing. Nor could it be disproved. The criterion of consistency requires that from the initial hypotheses we cannot derive contradictions. This holds for an empirical model as well as for a mathematical formal system.

Another requirement for a model is that "One must not introduce as many arbitrary constants as there are phenomena to be accounted for; they must establish connections among the various experimental facts and, above all, must lead to predictions." (Poincare, 1911) In studying human speech, we find that no two utterances are ever exactly the same. Yet, despite the wide differences, certain relationships appear to remain constant. Such constancy can be interpreted as signifying that these relationships are necessary. For example, when a speaker of a language repeats an utterance beginning with /b/, if an electrode is place on his lip over the orbicularis muscle, it will be found that this muscle contracts. The amount of muscle contraction will vary each time he makes a bilabial closure. When other speakers repeat the same utterance, it will likewise be found that this muscle is activated. One can thus establish a causal necessary relationship between the closure of the lips in a /b/ articulation and the contraction of the orbicularis muscle. One will find however, that the amount of muscle action varies. This may be due to contingencies representing essentially independent factors which may exist outside the scope of things that can be treated by the generalization under consideration, and which do not follow necessarily from anything that may be specified under its context. Such contingencies may include the amount of suction by which the electrode is being held onto the lip, or the growing fatigue of the subject. Having found some regularities which we provisionally suppose are the results of causal relations, we proceed to make hypotheses which would explain these regularities and permit us to understand their origin in a rational way. By explanation, of a given thing, we mean the demonstration that this thing follows necessarily from other things.

In observations and experiments, we must choose conditions in which the things we are interested in are not affected by chance. If the predictions based on our hypotheses are consistently verified in a wide range of conditions, and if, within the degree of approximation with which we are working, all failures of verification can be understood as the result of contingencies that it was not possible to avoid, then the hypothesis in question is provisionally accepted as an essentially correct one.

Even after correct hypotheses have been developed, the process does not stop there. For such hypotheses will, in general, lead to new observations and experiments, out of which may come the discovery of new empirical regularities, which in turn require new explanations, either in terms of a modification of existing hypotheses or in terms of a fundamental

revision of one or more hypotheses. Thus, theoretical explanations and models, and empirical verifications each complement and stimulate the other, and lead to a continual growth and evolution of science, both with regard to theory and with regard to practice and to experiment.

While this is commonplace knowledge in the physical and biological sciences, the relationship between theoretical constructs and empirical verifications has often been neglected in linguistics. The data of performance is not only necessary for the construction of models of performance and competence, but also necessary for the confirmation of both models, and possibly for the modification and revision of the models constructed.

The general requirements of a model of performance may be summarized as follows:

1) It must be based on the physical data of speech performance.
2) It must describe the phenomena of interest.
3) It must predict events, which predictions are confirmed by further experiments.
4) It must suggest certain causal relationships, singling out the necessary and sufficient elements, and excluding the contingent ones, thereby providing an explanation.
5) It must be consistent; it must not present contradictory explanations.
6) It must be testable; i.e. it must be disprovable.

Specific Requirements for a Model of Speech Performance

Before any useful model of performance can be constructed, further research is required which will provide more detailed knowledge about the encoding and decoding processes which take place. It seems logical, however, to start with the assumption that the semi-continuous acoustic signal that we call speech is the result of a number of discrete neuromuscular events. We cannot store in our brain motor commands for every utterance that we may wish to say, owing to the brain's finite storage capacity. Speech then is produced by the rearrangement of a limited number of stored items. One of the major problems to be resolved is the determination of the size and nature of these stored units. A second problem is to show how speakers encode a sequence of these discrete linguistic units into a continuously changing signal.

Our concern here is to suggest certain answers to these two questions. Because of the complexity of the problem, we shall direct ourselves only to that part of the entire communication process which is involved in the encoding of linguistic units into neural commands to muscles. We must therefore assume that at some previous level of brain activity some grammatical and semantic unit of sentence (i.e. noun phrase, noun, verb phrase, etc.) has been encoded into a series of discrete segments. These stretches of segments are then the messages which are to be encoded into signals or commands to activate the muscles in such a way that speech sounds are produced.

A simple diagram of this part of the system can be illustrated.

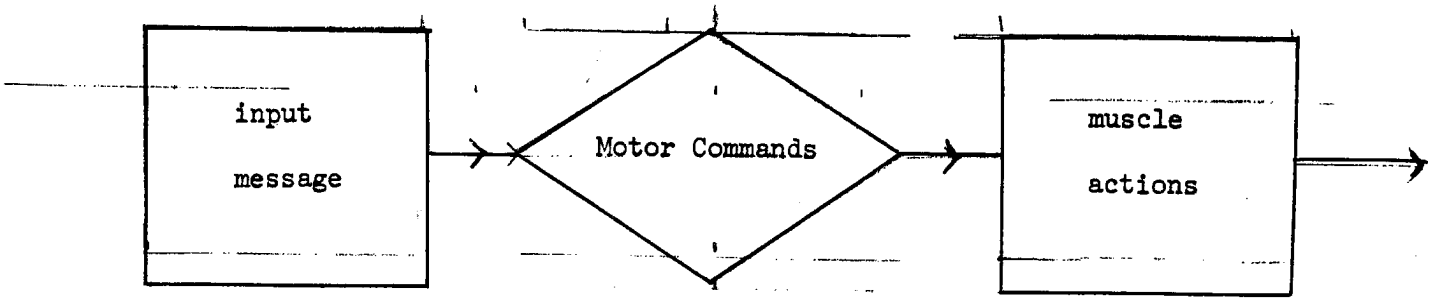


Figure 2

Even such a seemingly simple model presents complex problems. We are presented with a 'black box' situation, but one in which the only data we can examine is the output, i.e. the activity of the muscles, the articulatory gestures, and finally the acoustic signal. We are required to make certain assumptions about the input and the transformation of the input into the output which occurs in the Motor Command box. We are unable to look into the speaker's brain, and even if we did so, it is unlikely that this would reveal the encoding mechanism. While we know something about the afferent paths which lead from the receptor cells to the cortex and about the efferents leading to the motor areas used in speech, little is known about how and what the brain encodes and decodes, nor about the storage mechanism. Furthermore we have little knowledge of how the proprioceptive feedback mechanisms s involved in speech monitor the output to change the input. Besides, our interest is not in the physical workings of the neuro-muscular system, but in the transformation process which occurs.

What logical assumptions can be made about the input message?

First of all, we do not believe that on the performance level an entire sentence is generated, which sentence is then encoded into motor commands, as the Katz model seems to suggest. As already stated, too many discontinuities, false starts etc. occur in actual speech. According to Goldman-Eisler (1964) "Half of our speech time seems to issue in phrases not longer than three words." (119) Studies have shown that "hesitation pauses precede a sudden increase of information" (Ibid., 120), and that the part of a sentence surrounded by pauses are "clearly connected with articulation and must definitely be pronounced at one output." (Kozhevnikov, 1965). We would therefore support the proposal made by a group of Leningrad physiologists who investigated the time organization of articulation. They suggest that the linguistic unit which constitutes one message (one articulatory program) corresponds

to some semantic and syntactic unit smaller than a sentence, its average length amounting to approximately seven syllables. (Kozhevnikov, 1965)

This is not to suggest that there is stored in the brain a set of motor commands corresponding to each such message. It is even highly improbable (if not impossible) that there are individual sets of motor commands corresponding to the set of possible morphemes in the language. In English, for example, the number of possible morphemes (including morphemes of the size one to five syllables, and excluding all others) is far too large and may be of the order of 10^{22}. We are suggesting rather that the size of the message serving as input be a complex linguistic unit of a structure less than or equal to a sentence. And further that this unit is encoded into segments corresponding to phonemic segments.

We are assuming at the outset that these are phonemic segments since the 'reality' of the phoneme (as an individual element or as a bundle of distinctive features) is the accepted minimal linguistic unit of phonology. We shall, however, suggest an alternative hypothesis below, which makes a differentiation between the minimal unit in the competence model and that of the performance model.

A further assumption one can make about the input to the neuromuscular mechanism is that the phonemic segments into which the message has been previously encoded are bundles or sets of properties, or distinctive features (not necessarily those suggested by Jakobson or Halle), and that it is the individual features which are encoded into motor commands. This is suggested because of the complexity of the muscular movements involved in producing any single speech sound; the 10th, 7th, and 5th cranial nerves must send signals to the appropriate lip muscles, laryngeal muscles, tongue muscles. Since all speech sounds are produced by different combinations of a finite set of muscular movements, it is possible that the brain stores these subsets of commands rather than the complex set for each possible sound, and recombines them in certain prescribed ways. We can investigate the neuro-physiological correlates of these separate features, providing a physical basis for the concept of natural class of speech sounds, which hopefully will have some correlation with the linguistic grouping of sounds into natural classes. In other words, we can show that all sounds which have the feature 'bilabiality' are produced by action of the orbicularis muscle, all voiced sounds involve contraction of the laryngeal muscles, etc.

Superficially, this appears to be the simplest and most efficient storage mechanism, if it were indeed the case that the motor commands corresponding to the feature of bilabiality were invariant for all sounds. Our own experiments with electromyography showed this not to be the case; the duration and amount of muscle activity causing the closure for initial /b/ and /p/ were not the same. Additional rules changing the stored command would then be required for each sound containing this feature. One could conclude then that it might be indeed simpler to suggest that for each complex segment a set of motor commands be stored. The present state of knowledge can not resolve this point.

In either case, one need not assume that there is a mapping of phonemes or phonemic features onto individual motor commands. An intervening encoding process may occur which transforms phonemic features into phonetic features.

The model for speech production proposed by **Liberman and others** (1965) assumes that phonemic features are directly (without an intermediate stage) related to individual motor commands, as can be seen in Figure 3.

SCHEMA FOR PRODUCTION

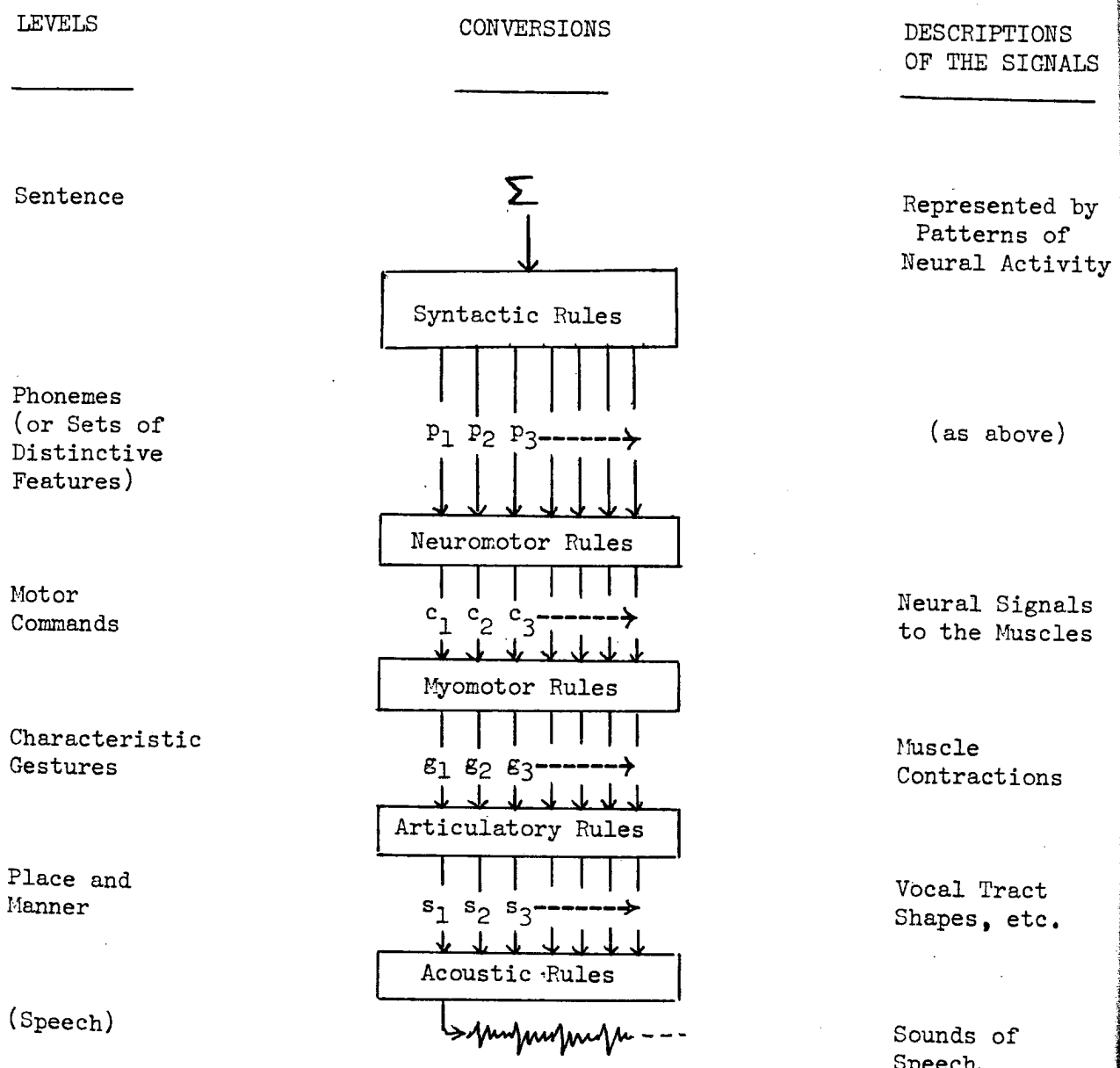


Figure 3

We have already pointed out above that we do not believe the input to the neuromotor rules box to be a sentence as is suggested by the Haskins model. Does the rest of their model comform to the observable data? According to them the process starts "with the message in the form of a phoneme sequence (alternatively, sequential sets of distinctive features) taking this sequence as the input to successive converters that operate by neuromotor rules, myomotor rules, articulatory rules" etc. (1.11)

> In the first of these operation, the neuromotor rules serve to convert the ordered string of phonemes into a temporal sequence of neural signals to the muscles of articulation." (1.11) (my emphasis)

They do suggest that the temporal overlapping of incoming instructions to the muscles complicate matters so that

> the shapes that result will no longer stand in one-to-one correspondence with the phonemes, but will reflect at each instant the interacting influences of several phonemes.... We can summarize the effects of the conversion from contraction to shape by saying that it is complex at best and almost always introduces an encoding of the sequential input units into output units of about syllabic size. (1.12) (my emphasis)

The encoding that takes place according to the model is, then, from phoneme to motor command; the overlap of motor commands influences the contractions and shapes taking place so that the articulations resulting (the output) are syllabic in structure.

Our own investigations reveal certain data not easily explained by this model. We do agree that the features are not invariant, that they are dependent on the context but we are not sure that the Haskins' suggestion as to how the variation occurs can explain all the phenomena.

Since a performance model must predict and explain the kind of variation which occurs, we shall present a summary of some of our findings regarding the activity of the orbicularis oris in the articulation of CVC utterances, in which C = /b,d/ and V = /i,ʊ,u/. We shall attempt to see whether the Haskins Model presents the best possible explanation for our observations.

1. In utterances in which both C_1 and C_2 = /d/, e.g. /d_d/, no muscle activity was recorded when V = /i/; muscle action potentials were recorded when V = /u/ and /ʊ/.

2. In all utterances in which /b/ occurred, muscle activity was recorded; for initial /b/ the action potentials occurred prior to the onset of the audio signal, for final /b/ the action potentials (AP's) occurred after the onset of the audio signal.

3. When /b/ occurred initially, the muscle action was greater in amplitude and duration than when /b/ occurred finally.

4. The muscle action associated with an initial and final /b/ was

unaffected by the quality of the vowel which preceded or followed it.

5. The action potentials produced in the articulation of /dud/ i.e. of the vowel /u/, started prior to the onset of audio and extended after the onset of audio. There was less muscle action after the onset of audio when /u/ was preceded by /b/ than when /u/ was preceded by /d/.

6. The muscle activity associated with /u/ and /ʉ/ was independent of the consonant which followed, i.e. when followed by a /b/ which also showed action potentials there was no difference than when followed by a /d/ which showed no muscle action.

7. The muscle action potentials produced in the utterance /bʉd/ were relatively indistinguishable from that for /bid/.

This may be generalized as follows:

In a C_1VC_2 utterance, let C_1 = segment 1; V = segment 2; C_2 = segment 3.

Let M = motor command to muscle motor units. (then M1 = motor command for segment 1, M2 = motor command for segment 2, etc.)

Let $t_i(M)$ = moment of time of issuance of command. (We shall disregard for now the delay time between the issuance of the command and the execution of the response.) then $t_i(M_1)$ = moment of time when command is issued and muscle contracts.

Let $t_f(M)$ = moment of time muscle action ceases.

Let d(M) = duration of the muscle action resulting from M.

And a(M) = amount of muscle action resulting from M (the frequency of discharge rate plus the number of motor units innervated.

Finally, let $d(t(M_x, M_y)$ = duration of the time period from $t(M_x)$ to $t(M_y)$.

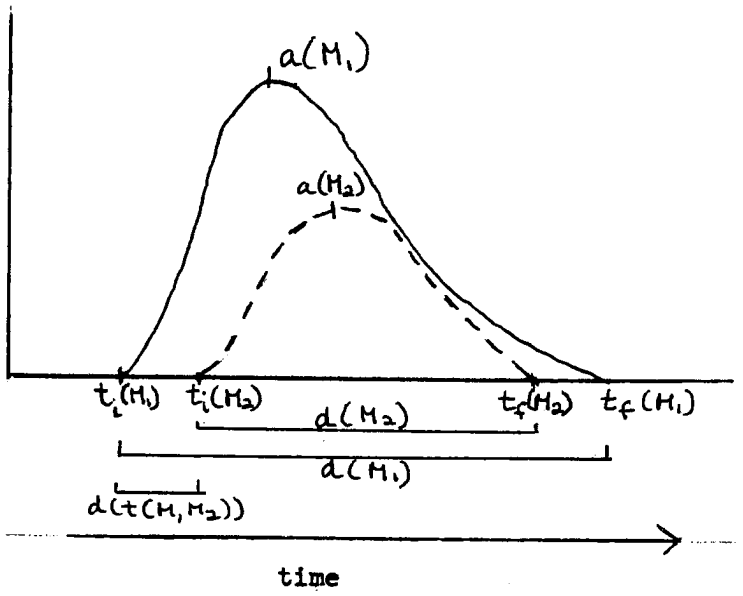


Figure 4.

We can make the following generalizations from the data:

1) If $t_i(M_2) < t_f(M_1)$ and if $a(M_2) \leq a(M_1)$, then $a(M_1) + a(M_2) = a(M_1)$.

Two motor commands directed to the same muscle do not result in a simple addition of the number of motor units contracting. (We are of course assuming here that there are separate commands for the individual segments, rather than one set of commands corresponding to more than one segment.) The second gesture seems to be 'locked out' by the first. The result resembles that category of devices which upon being triggered into action remain insensitive to further stimuli during their period of firing and recovery. In other words if n motor units are recruited to produce a certain gesture, and if during this gesture, a command is issued for m motor units to contract and if $m \leq n$, the set of m motor units will be a subset of the set of n motor units.

2) If $d(M_2) + d(t(M_1, M_2)) \leq d(M_1)$, then $d(M_1) + d(M_2) = d(M_1) = d(M_1) \cup d(M_2)$.

If a second gesture occurs within the duration of the first gesture, the total duration of both gestures is that of the first.

3) If $d(M_2) + d(t(M_1,M_2)) > d(M_1)$, then $d(M_1) + d(M_2) = d(M_1) + d(M_2) - d(t(M_1,M_2))$.

The duration of the muscle activity does not decrease with two gestures.

4) If $d(M_2) + d(t(M_1,M_2)) > d(M_1)$, and if $a(M_1) > a(M_2)$, then, $a(M_2)$ is decreased.

If a second gesture is preceded by a stronger gesture, the second gesture shows less muscular activity. Physiologically this may be explained by the fact that a strong effort is required to initiate the closure of the lips with many motor units recruited for the task and with the frequency of discharge increasing until the lip closure occurs. Less effort is needed to sustain the partial closure.

The above statements are generalized from only a single muscle and a few utterances. However, they represent a possible hypothesis to be tested for other muscles and other utterances. It may be added that they do hold for this muscle for 12 American English vowels and for /p/ as well as /b/. And they are in keeping with general physiological facts. "...The number of excited motor neurons may change as a function of the original state of the motor neurons." (Kozhevnikov, 1965)

Using the Haskins model to explain the above phenomena, one must assume different sequences of phonemes, e.g. /bib/,/bud/, /dib/, etc., as input to the Neuromotor Rules mechanism. A further assumption is that there is one neuromotor rule (or set of such) corresponding to each of the phonemes serving as input. The overlapping instructions for neighboring phonemes influence the articulatory shapes which result, so that, for example, the gesture for an initial /b/ will differ from that of a final /b/. What occurs in fact is that the commands to the muscles must be different, since the muscle action potentials recorded for initial and final phonemes differed. While the resulting vocal tract shapes may differ, these

are due to different commands. And further, the differences between the initial and final gestures are not due to the neighboring sounds, but to the phonological context, i.e. the ordered position of the phoneme in the linguistic unit. The data reveals an invariance in the muscle action connected with initial /b/ and with final /b/ but definite differences between them. There is no way to explain this difference by reference to overlapping instructions to the muscles. In this sense, the Haskins model is inadequate.

Alternative suggestions are possible. We can assume, instead, that the input to the neuromotor mechanisms is not a sequence of phonemic segments, but of phonetic segments (allophones). Each phonemic feature is then encoded into a phonetic feature, which in turn is encoded into a discrete motor command. This may be represented as follows:

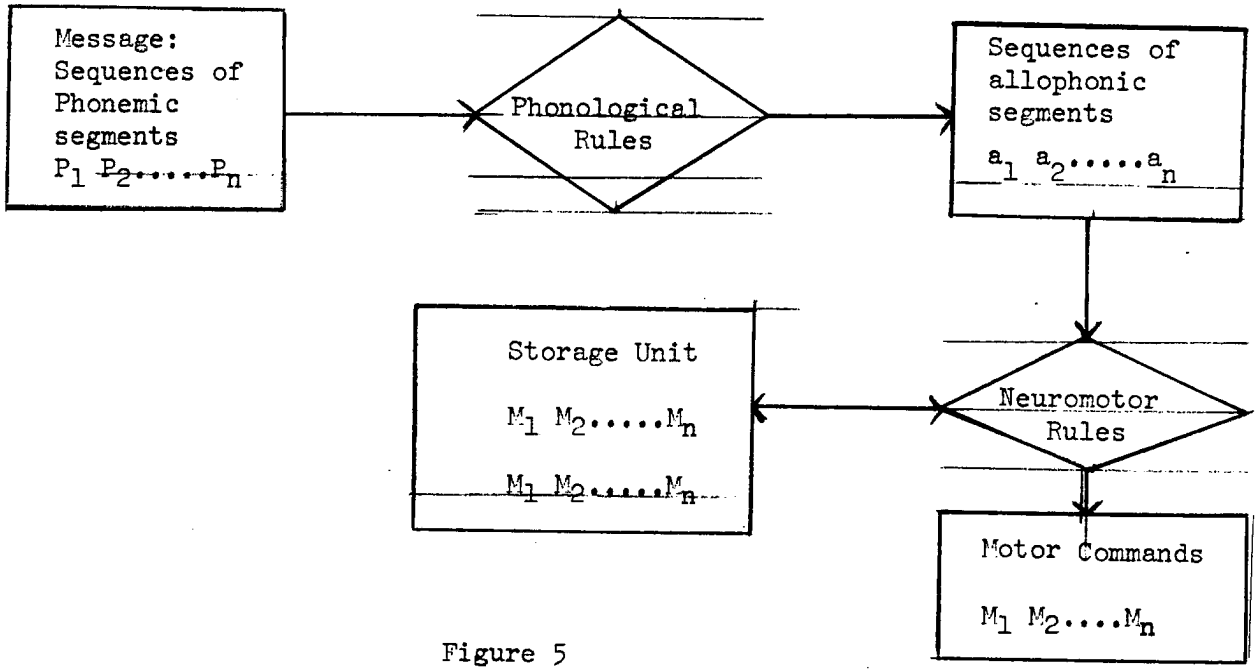


Figure 5

Such a system assumes that there is stored in the brain a separate command for every possible allophone. Using this model, one would have to recognize allophonic differences which seem to be the result of overlapping sets of motor commands as well as those which are not. This would be a highly inefficient method of encoding.

A more plausible hypothesis is one which assumes a distinction between what Ladefoged has called extrinsic and intrinsic allophones (**Ladefoged**, 1966). For our purposes we shall define extrinsic allophones as those which are not due to coarticulation factors, but which require a different set of motor commands, representing either a change or addition of features, (e.g. the **aspiration** of an initial stop) or change in value (duration of muscle contractions or number of motor units recruited) of the commands issued, (e.g. different gestures for #b___ and __b#).

Intrinsic **allophones** can be processed and transmitted directly as signals to the muscles. Thus, the decrease in the muscular activity for an /u/ preceded by a /b/ would be the result of the previous muscular

state.

This model could account for anticipatory phenomena such as that occurring in vowel harmony languages, in which features of preceding vowels are determined by the presence of features of subsequent vowels. In Twi, for example, the quality of the vowel in a prefixed personal pronoun is determined by the vowel in the verb stem; [mɪdɪ] 'I cause' as opposed to [midi] 'I eat'. The quality of the first vowel is obviously not the result of immediate overlapping instructions. Such vowel harmony languages also provide additional justification for the assumption that the message which corresponds to one articulatory program is larger than the single morpheme, since such phenomena occur over stretches involving more than one morpheme.

The model as proposed is still deficient in that it does not reveal how intrinsic allophones are processed, i.e. how the phonetic context of neighboring sounds interferes or influences antecedent or subsequent sounds. Referring to overlapping instructions does not provide an explanation for the data reported above. The variations which do occur can be the result of at least two possible processes.

The first assumes that the phonemic segments are transformed into extrinsic allophones. By a feedback mechanism, the allophones which serve as input to the motor-command encoding devise are monitored and changed depending upon what state the muscles are in when the command for the new segment is issued.

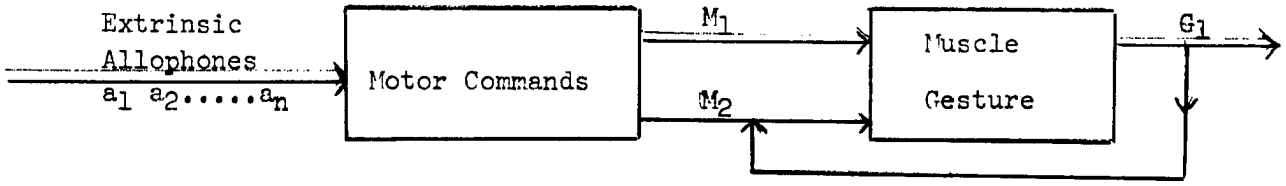


Figure 6

It appears however, that the rapidity of the articulation process coupled with the delay time necessary for the proprioceptive feedback would disallow feedback from the peripheral organs to modify the commands from the nervous system in time. However, in certain systems, the output of the system may be monitored as part of the command itself. (Dr. Moore, private communication) i.e. the nervous system can look at the command rather than at the output, and change a subsequent command in keeping with the requirements to achieve the new gesture.

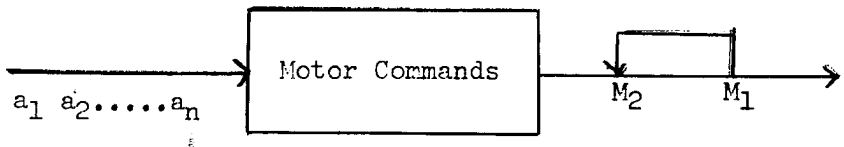


Figure 7

This is one possible hypothesis to explain the processing of intrinsic allophonic features. It would account for the rules stated above in the following way. A series of neural commands are sent to a given set of motor units of the orbicularis muscle to contract for the articulation of /b/. The nervous system monitors the commands sent; it then sends this knowledge back to the neural center and compares the next command for an /u/ with the projected muscle response of the /b/, modifying the command for an /u/ in keeping with the state the muscle is in.

An alternative hypothesis can be suggested, namely that the linguistic unit corresponding to individual sets of motor commands is not of the size of a phoneme or phonetic segment but rather that of a syllable, or some combination of phonemic segments. Much supporting evidence for this position has been advanced. For example, in a CV syllable, in which the vowel is marked by a rounding feature (or protrusion), the protrusion occurs simultaneously with the articulation of the consonant, or in some cases even before. Other such co-articulation phenomena are documented in the literature. (Ladefoged 1957, Fujimura 1961, Ohman 1964, Kozhevnikov 1965, et alia).

Fry (1964) supports such a position:

> ...it is at least plausible that...syllabification is one feature of the brain's control of motor speech activity and that the true function of the syllable... is to form the unit of neural organization. This would mean that during the production of speech the brain mechanism responsible for timing the action of all the muscles used in breathing, phonation and articulation arranges the time scheme for a complete syllable as a unit and the operating instructions are then fed forward to the muscles in accordance with this scheme whilst the timing for the syllabic unit is being organized. (219)

Kozhevnikov and Chistovich present evidence suggesting that syllable commands are rhythmically organized by a separate rhythm generator in the nervous system, distinct from the articulatory generator. (115)

If this is indeed the case, either one must conclude that a many to one encoding process occurs prior to the mechanism which transmits the motor commands, or that the message to be transmitted is originally encoded into units of the size of a syllable rather than smaller phonemic segments.

It is possible that the motor commands corresponding to the syllables are stored, i.e. for every possible syllable in the language a separate set of motor commands is stored. This would then cancel the necessity for any feedback mechanism to monitor and change the signals corresponding to individual phonetic segements. The transitions between syllables might require such feedback. It is more probable that the syllable units are composed of segments which are programmed according to the feedback mechanism suggested above. This would require much less storage capacity. In the first instance there must be stored articulatory commands for approximately 200,000 syllables in English. In the second instance there need be stored only commands for the extrinsic allophones with

intra-syllabic rules governing the serial ordering. The set of such commands would be on the order of a few hundred.

Another justification for the second suggestion is suggested by the Leningrad group. Their investigations reveal that movements required by consonants are accomplished consecutively within a syllable. While this may not be true of all consonants in all languages, it must be true for the sequences of consonants or consonant-vowel in which antagonistic muscular movements follow each other. Consecutive ordering of consonants and vowels implies units smaller than the syllable, which units would correspond to phonemes, or extrinsic allophones.

In examining the process which encodes one message into motor commands, another phenomenon which occurs in speech raises an interesting question. This deals with spoonerisms and slips of the tongue which produce errors such as the substitution of "ship slod" for "slipshod." The occurrence of such errors provides another justification for the assumption that prior to articulation there must be stored sequences of segments which correspond to articulatory program. The transposition of the speech sounds may occur at a level of brain activity prior to the motor command encoding, i.e. when the word is stored prior to the encoding process the segments are rearranged. One can justify such an assumption by the fact that while one may say 'shlip sod', 'slid shop', 'shipslod', or 'shopslid', one would not err by saying 'psiladsh.' This shows that the substituted utterance obeys the morpheme structure and phonological rules of the language, which must function prior to the neuro-muscular encoding process.

On the other hand, 'dip slosh' [ɪipslaš] is a possible sound sequence in the language, but unlikely as an error in this instance. The suggestion that the confusion is a result of a transposition of phoneme segments does not explain why 'dip slosh' is a highly improbable error for 'slipshod'.

The problem is not solved by a model which would provide for a transposition after the segments are encoded into syllables if these syllables are not composed of smaller segments, since it is parts of syllables which are transposed or anticipated, not whole syllables.

A possible explanation suggests that the confusion occurs after the message is encoded into syllables, which are segmented into extrinsic allophones. A sequence of such syllables thus constitutes one articulatory program. The segments constituting each syllable must have sequential ordering, so that only initial consonants or clusters, vowels, and final consonants may interchange if and only if the transpositions are in keeping with the phonological rules of the language. In other words 'Shlitz beer' will not produce 'shbitz leer,' but 'fly paper' could produce 'ply faper.' This suggests that order is a property of each segment at all points along the tranmission system. This explanation would support Fry's observation that "such errors are practically never corrected until a whole syllable at least has been emitted." (219)

One may then suggest the following (obviously oversimplified) alternative models for the speech encoding system.

The errors of perseverance or anticipation would have to occur in Model A either between boxes 2 and 3 or between boxes 3 and 4, in which case the sequences would have to rerun through the P rules. Model B is a more complicated system requiring a syllable encoder mechanism, not required in Model A. The errors must occur after the sequences of allophones have been separated into sets corresponding to syllables due to the reasons cited above, i.e. only syllable initial sounds are substituted for each other, etc. But since we do not transpose sounds producing combinations incompatible with the structures of the particular language the phonological rules will have to apply after box 6. This appears to be needlessly inefficient.

Model A may be further modified by omitting Box 2 and assuming that the syllabic units are encoded directly as extrinsic allophones:

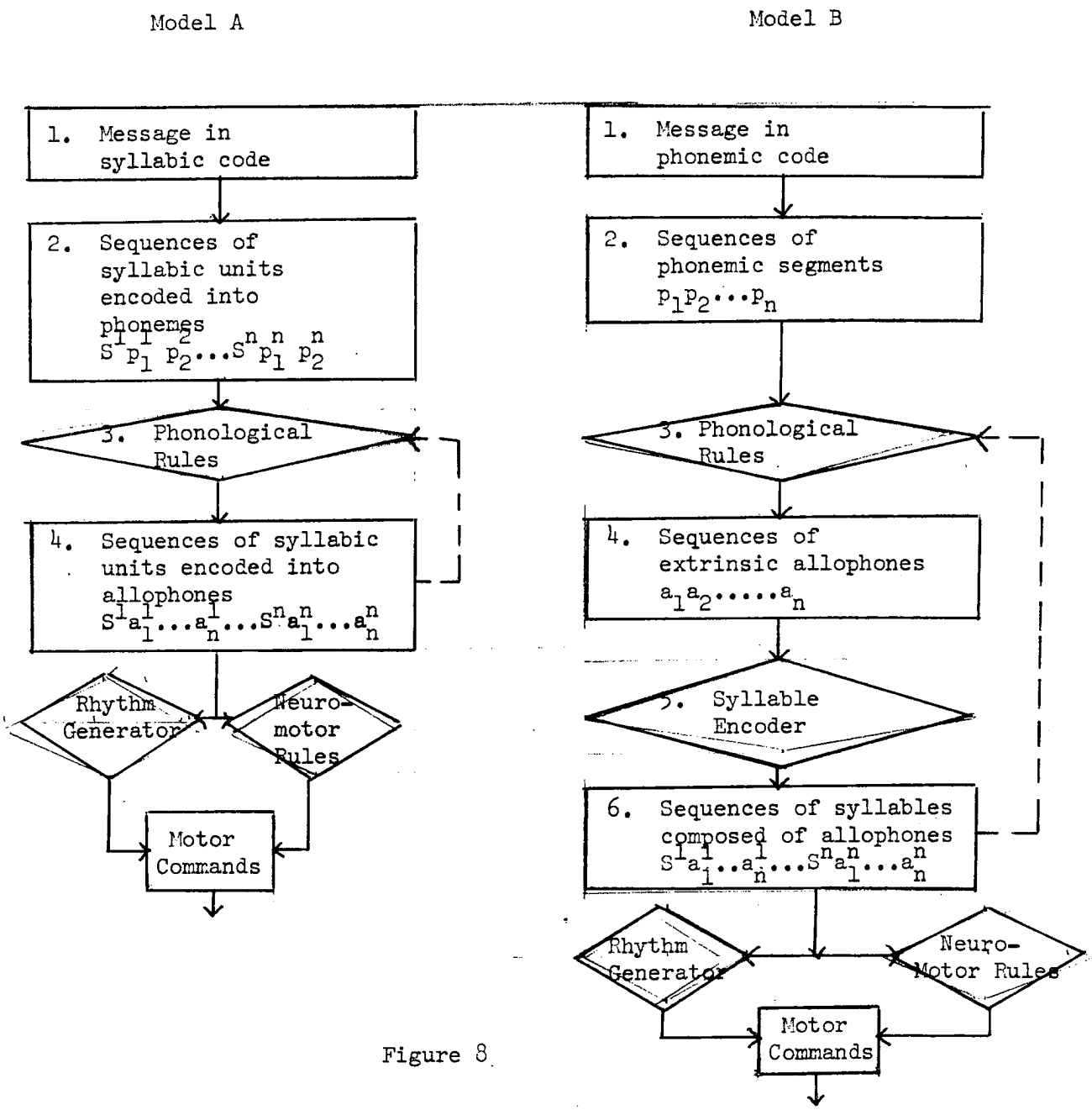


Figure 8

Model C

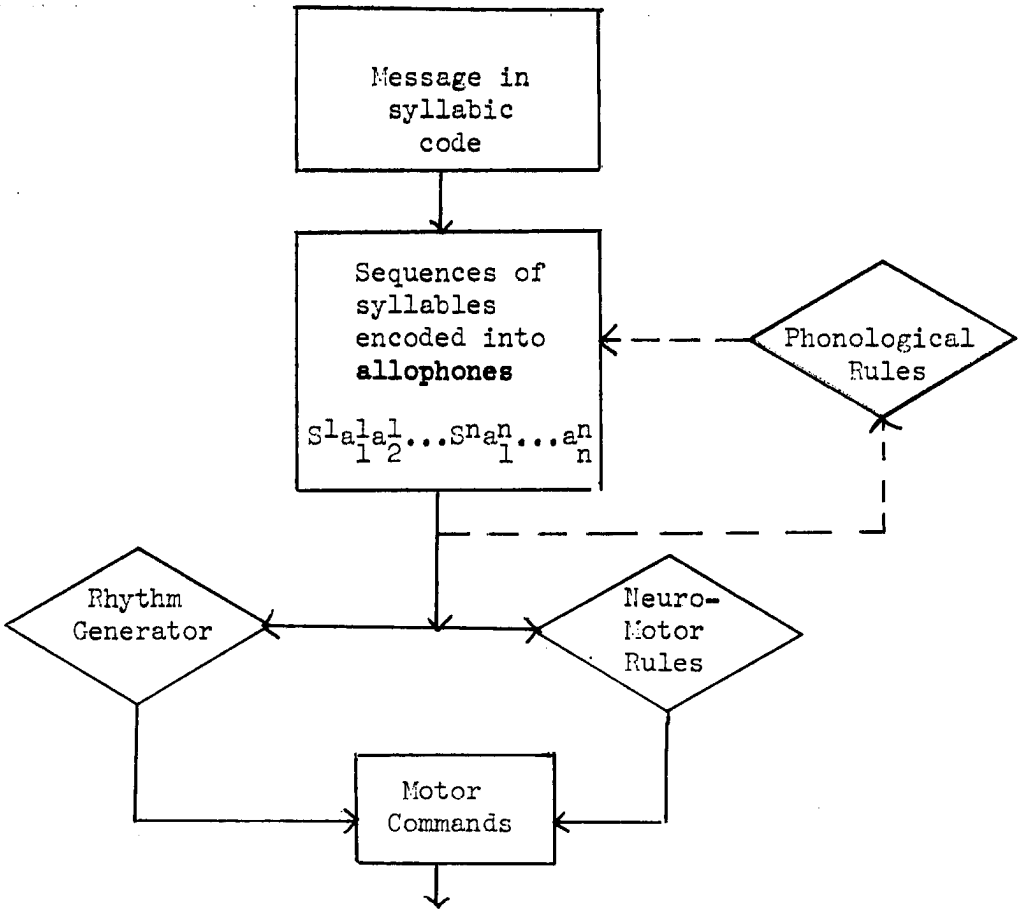


Figure 9

If this occurs, and if we accept the 'psychological reality' of phonemes we are saying that the encoding process does not include any phonemic encoding and is different quantitatively from the decoding, perception, process.

While hearing and understanding may not be the reverse decoding process of the speaker's encoding, there is still too little known about each stage of the system to assume either to be the case. Halle and Stevens (1964) have suggested a generalized model for both speech production and recognition. They were mainly concerned with an analysis by synthesis model "in which mapping from signal to message space is accomplished largely through an active or feedback process." (604) The speech production aspect is merely mentioned in general outline as to what might occur. The Haskins group have proposed as well as the model outlined above a speaker-listener model. But their principal interest has been directed toward the problem of how a listener can recover phoneme segments from a semi-continuous acoustic signal, they propose as an explnation a motor theory of speech perception. Only in so far as this model presupposes the speech production system have our comments been relevant.

Halle and Stevens (1964) have pointed out that "further research is necessary before the remaining components can be realized and before the system can be designed to function as a whole." (611) They set down as a major task the need to "establish the generative rules describing the conversion of phonetic parameters to time-varying speech spectra. Among such rules, they point out that the least understood to date are those showing the relations between "the phonetic parameters and the vocal-tract geometry and excitation characteristics," (611) and correctly suggest that "this work must...take physiological factors into consideration more directly, through the use of cineradiography, electromyography, and other techniques." (612)

We are in agreement with this observation. In this light, the work of the group of physiologists in Leningrad (Kozhevnikov, et al 1965) is an important contribution to our understanding of one of the subcomponents. Their work attempted to determine the effect of the rate of speech on the relative duration of words, syllables, and speech sounds, as part of their investigation of the temporal organization of articulation. They also treat the temporal ordering of motor commands within a syllable.

We view our own research using electromyography, some of the findings which are presented above, as merely additional data which must be considered in the construction of an adequate model of performance. New data, and the discovery of new relationships, will no doubt alter drastically the models suggested above. We thus view the presented schemes as working hypotheses, on which to base further experiments.

Bibliography

Chomsky, N. (1965) Aspects of the Theory of Syntax, Cambridge, Mass.: The MIT Press.

Fry, D.B. (1964) "The functions of the syllable" Z. Phon. 17, 215-237.

Fujimura, O. (1961) "Bilabial stop and nasal consonants: a motion picture study and its acoustical implications" J.S.H.R., 4.3 223-247.

Goldman-Eisler, Frieda (1964) "Discussion and further comments" New Directions in the Study of Language, ed. by E.H. Lenneberg, Cambridge, Mass.: The MIT Press, 109-131.

Halle, M. and Stevens, K. (1964) "Speech recognition: a model and a program for research" The Structure of Language, ed. by J.A. Fodor and J.J. Katz, Englewood Cliffs, N.J.: Prentice-Hall, Inc. 604-612.

Katz, J. and Postal, P. (1964) An Integrated Theory of Linguistic Descriptions, MIT, Research Monograph No. 26, Cambridge, Mass.

Katz, J. (1964) "Mentalism in linguistics" Lang. 40, 124-138.

Kozhevnikov, V.A. et al. (1965) Speech: Articulation and Perception. Washington, D.C.: U.S. Dept. of Commerce - Joint Public Research Service.

Kuhn, T.S. (1962) The Structure of Scientific Revolution, Univ. of Chicago Press.

Ladefoged, P. and Broadbent, D.E. (1957) "Information conveyed by vowels" J. Acous. Soc. Amer., 29:1, 98-104.

Ladefoged, P. (1966) The Nature of General Phonetic Theories, Georgetown Univ., Monograph No. 18, Languages and Linguistics.

Liberman, A.M.; Cooper, F.S.; Studdert-Kennedy, M.; Harris, Katherine S.; and Shankweiler, D.P. (1965) "Some observations on the efficiency of speech sounds" Status Report on Speech Research, SR-4, New York: Haskins Labs.

Ohman, S.E.G. (1964) "Numerical model for coarticulation, using a computer simulated vocal tract" J. Acous. Soc. Amer.36.5, p. 1038. (abstract of paper delivered at 67th meeting, ASA)

Platt, J.R. (1962) The Excitement of Science, Boston: Houghton Mifflin.

Platt, J.R. (1964) "Strong inference" Science 146, 3642, 347-353.

Polya, G. (1954) Mathematics and Plausible Reasoning, Princeton, N.J.: Princeton Univ. Press.

Popper, K.R. (1959) The Logic of Scientific Discovery, New York: Basic Books.

Saporta, S. (1965) "Review of Psychology, Study of a Science: Study II. Edited by Sigmund Koch, Vol. 6" Language, 41, 1, 95-100.

A New Photo-Electric Glottograph

John Ohala

This is a report of our work on a device which is being developed for use in research on laryngeal activity during speech. The device is conceptually similar to the apparatus used by Sonesson (1959, 1960) in his important work on the vibratory pattern of the vocal folds (Cf. Zemlin, 1959). We shall follow Sonesson and call this device a photo-electric (diaphanoscopic) glottograph. (Cf. the terminology of Fabre (1957, 1958, 1959) who calls his apparatus a glottographe de haute frequence.)

Sonesson's apparatus consisted of a strong DC light source place against the neck just below the larynx, the light from which suffused the subglottal region and transilluminated the glottis. A curved inflexible Lucite "light pipe" was passed through the mouth and into the throat and terminated at about the level of the epiglottis. It was aimed at the glottis and transmitted the light coming through the glottal chink to a light sensor outside the mouth. The light sensor transduced the variations in light intensity into variations in electrical voltage, thus providing a direct measure of the degree of opening and closing of the glottis. These variations in electrical voltage could be recorded in a variety of standard ways and then analyzed. Sonesson limited his investigations to steady-state vowels produced with varying pitch and loudness. Due to the rigid light pipe in the subject's mouth, studies of vocal cord vibrations during connected speech would have been extremely difficult if not wholly impossible.

An improvement on Sonesson's technique which immediately suggested itself, then, was inserting the photo-electric pick-up into the throat via the nose so that it would not interfere with the normal speech articulations of the tongue and lips, thus permitting an examination of the vocal folds' activity during connected speech. Thanks to the extreme miniaturization of electronic components these days, this is easily accomplished.

In a flexible transparent plastic catheter (4mm. O.D.), sealed at one end, a small photo-resistive diode was encased so that it was about 25 cm from the sealed end. (The light sensor was a Texas Instruments LS-400; 2mm diameter; 13mm long.) The leads from the light sensor extended through the catheter and emerged out the open end. The catheter was inserted into the throat and esophagus via the nasal passage so that the light sensor was about 15-16 cm from the external nares. The extra length of catheter preceding the light sensor and extending into the esophagus not only helps to stabilize the position of the light sensor relative to the glottis, but also adds immeasureably to the comfort of the subject, since a loose tube dangling in the pharynx tends to trigger a gag reflex. With this feature plus the catheter's pliancy and thinness we obtained a device that caused no discomfort to the subjects and after an initial 5 minute period following insertion caused no disruption of the natural speech articulations. Tape recordings of subjects speaking with and without the catheter in their throats were indistinguishable.

An approach similar to Sonesson's was used by members of the Haskins group before 1961 (Cooper 1964). They reversed the position

of light source and photo-electric pick-up relative to the glottis, having the light transmitted to the supraglottal laryngeal cavity by a thin fiber optic inserted through the mouth, the light variations being read by a photo-cell placed against the trachea just below the larynx. As early as 1964 they mentioned the desirability of inserting the fiber optic through the nasal passage (Lisker and Abramson, 1964) and achieved this in 1966 (Abramson, 1965; Lisker, et al., 1966). A full description of the Haskins' apparatus has not yet been published. They do report some difficulty in obtaining a comfortable velo-pharyngeal closure.

The circuitry we used is very simple: a 6v power supply, a resistor and the light sensor are wired in series, as shown in Figure A. The manufacturer indicates the light sensor's frequency response as flat ± 2 dB. from DC to 90 kHz., with a 1000 ohm resistor. We found that by increasing the resistance we could enhance the light sensor's sensitivity at the expense of the frequency response. We therefore introduced a switch with which we can choose one of two values of resistance depending on which characteristic of the sensor we are interested in most at the time.

At maximum sensitivity, using a 68K ohm resistor, the frequency response is flat ± 2dB. from DC to 1200 Hz; with a 6.8K ohm resistor, the frequency response is extended to 6000 Hz, with an approximately 10 dB drop in sensitivity relative to that obtained with the larger resistance. The only other aspect of the electrical system worth mentioning is that due to the relatively weak signals obtained (with our apparatus, seldom more than 10 mv.) all external leads must be shielded and the subject grounded.

The light source is a 100 watt incandescent lamp in an ordinary spotlight housing with a 3" plano-convex lens; it is powered by a 120 v DC power supply. The spotlight does not generate an uncomfortable amount of heat upon the skin. Rubbing the throat beforehand with glycerine or Vaseline seems to be adequate for the subject's comfort. In order that the position of the glottis remain in as constant a position as possible relative to the spot on the neck where the light is shining, the subject's head and the light must both be fixed. Even with this precaution, though, the larynx still moves somewhat; changes in pitch, changes in the pressure differential across the glottis, changes in tongue position, etc., all involve movements of the larynx. There is also some movement of the light sensor relative to the glottis. This is due mainly to the action of the velum upon which the catheter rests. We know of no simple way of eliminating this. For the most part it is not a serious problem, merely registering the glottal openings with decreased amplitude.

The types of records produced by the glottograph are illustrated in Figures 1-7. All were made of one subject, a 24 year old male speaker of the Midwestern dialect of American English. A glottogram is a measure of glottal area as a function of time, evolving from left to right. Thus a positive-sloped line represents an opening movement of the vocal cords, and a negative-sloped line, a closing movement. In these figures there is no absolute calibration of the amplitude of the glottograms, i.e., we have not attempted to quantify any given point on the ordinate with respect to a particular glottal area. However, in any given glottogram, the relative amplitude levels are consistent.

The glottogram on the cover and Figures 1 through 6 were

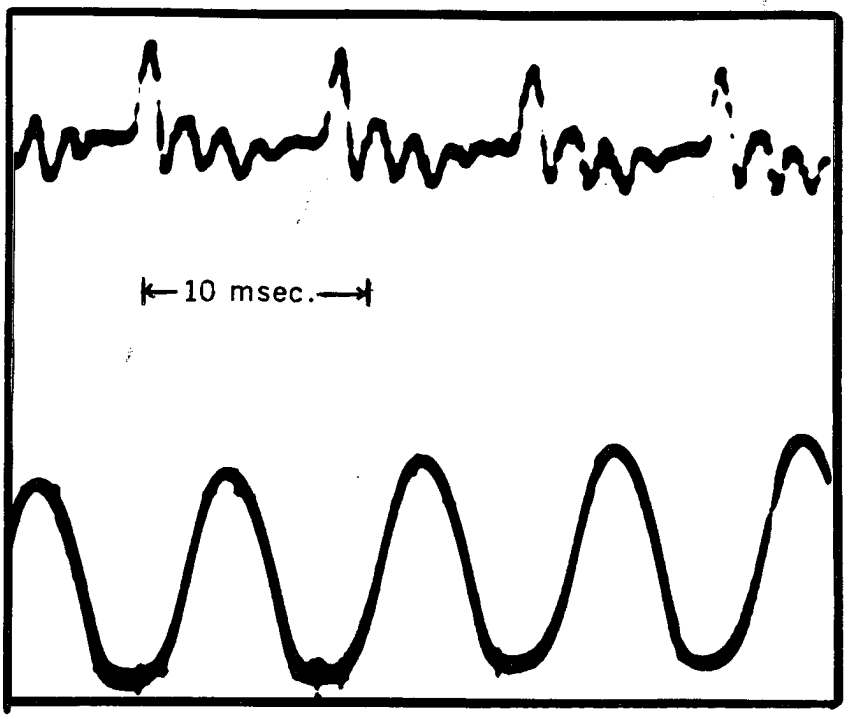


Figure 1. Glottogram (bottom) and simultaneous microphone signal (top) of vowel [ɑ] produced with breathy voice; F_o = 122 Hz.

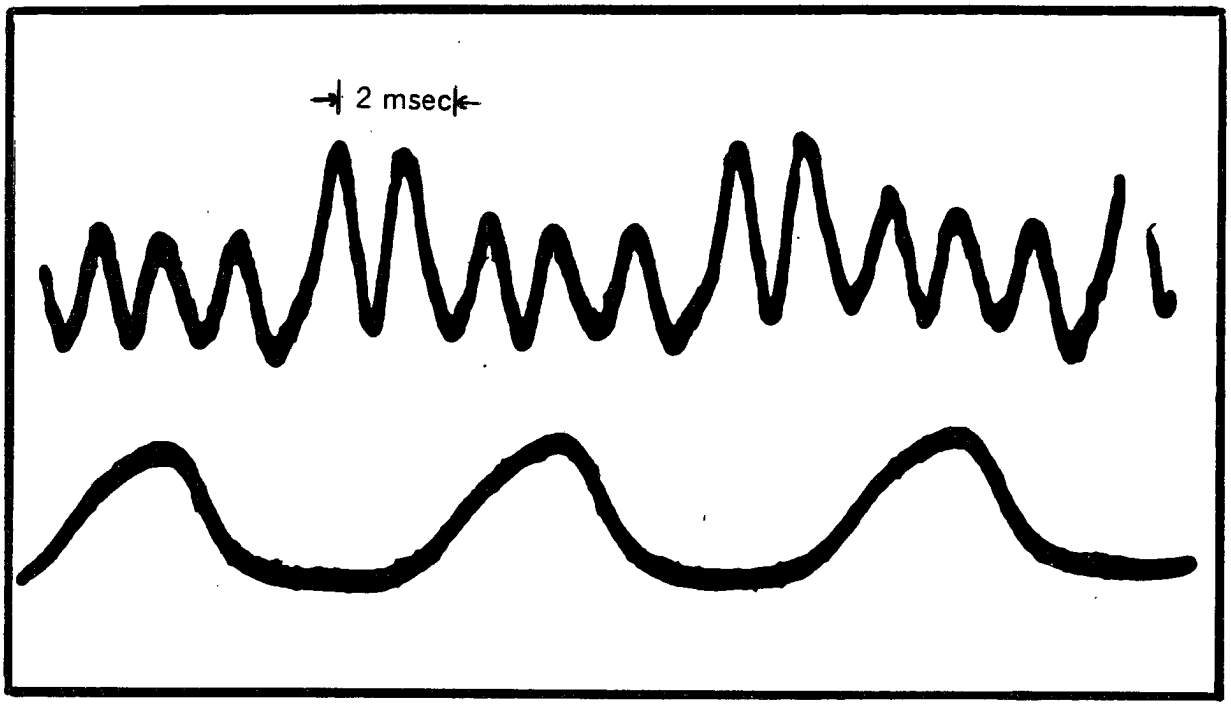


Figure 2. Glottogram (bottom) and microphone signal (top) of vowel [ɑ] produced with normal voice; F_o = 142 Hz.

photographed from the screen of an oscilloscope. Figure 7 is a **spectrogram** made of a glottographic signal tape recorded on an ordinary tape recorder whose frequency response was flat ± 2 dB down to 50 Hz; the upper frequency response of the system was limited by the light sensor, not the tape recorder, at about 1500 Hz. The simultaneous speech sound wave which accompanies each glottogram was picked up by a microphone placed 10-15 cm in front of the mouth. The upper part of Figure 7 is a spectrogram of speech recorded by such a microphone.

A comparison of Figures 1 and 2, both of the vowel [ɑ], the first one produced with breathy voice, and the second with normal voice, shows an obvious decrease and even disappearance of the closed period between each successive glottal opening as one goes from normal voice to breathy voice. Although not evident from the figures because of the lack of vertical calibration, the glottal openings during breathy voice were larger (roughly 5 times the glottal openings for normal voice) and, furthermore, start from a higher "base line"--indicating that the vocal cords do not close completely. Breathy voice has been described elsewhere as involving an incomplete closure of the vocal cords (Moore 1962, Catford 1964, Ladefoged 1964).

The general shape of the glottographic pulses during sustained phonation may vary between being roughly symmetrical to being quite asymmetrical, with the closing phase being shorter and more abrupt than the opening phase. This latter shape of the glottal pulse, quite typical for normal voice, has been reported in or illustrated in Sonesson (1960), Cederlund, et al. (1960), Fant & Sonesson (1962), Miller (1959), Holmes (1962), and Lindqvist (1965). (See below for justification of equating glottographic pulse to glottal volume velocity wave form.) Sonesson has shown that the degree of symmetry of the glottal pulse is strongly correlated with the relative intensity of voice.

It can also be seen from Figures 1-6 that the maxima in the speech sound wave occur during minima in the glottal area curve. This relation is more ambiguous in high vowels because the frequency of the first formant is close to the fundamental frequency, but it is still probably correct. This agrees with findings of Fant & Sonesson (1962) and Cederlund et al. (1960). Holmes (1962) in his studies of the glottal volume velocity waveform obtained by inverse filtering found that while formant 1 was excited only once per glottal opening, formants 2 and 3 could be excited twice per glottal opening, once at the closure of the glottis and once at the opening. Figure 3, of the vowel [i], seems to show this phenomenon where the formant 2 has two amplitude maxima at moments A and B, corresponding to the opening and the closing of the glottis.

Finally, something quite obvious, any glottogram that is time calibrated (by the scale on the face of an oscilloscope--obliterated in these reproductions--or by a simultaneous sine wave of known frequency) can quite simply and unambiguously give a measure of the frequency of voice, not only during vowels, but during nasals and voiced fricatives as well. It should be a trivial point to note that the frequency of the glottal vibrations equals the frequency of the sound emitted from the mouth; however Fabre (1957) shows some confusion about this. Noting that when his subject attempted to match his voice frequency to that of a pure tone, the frequency of his vocal cord vibrations was one

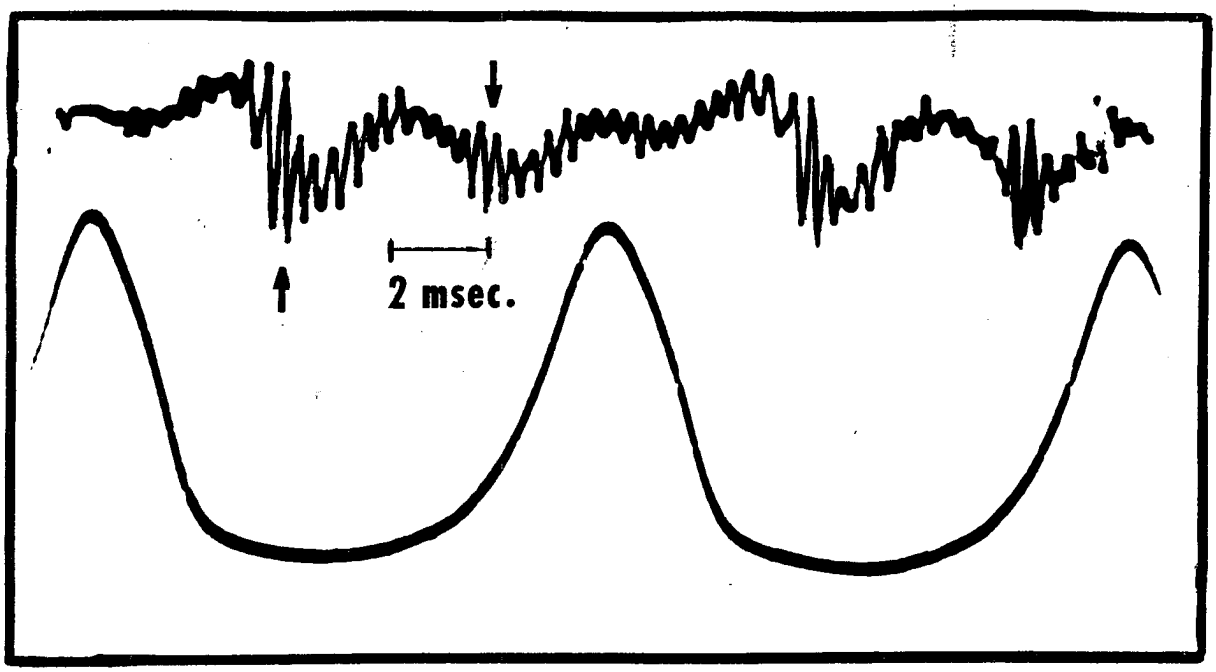


Figure 3. (Retouched) Glottogram (bottom) and microphone signal of vowel [i]; normal voice; F_o = 94 Hz.

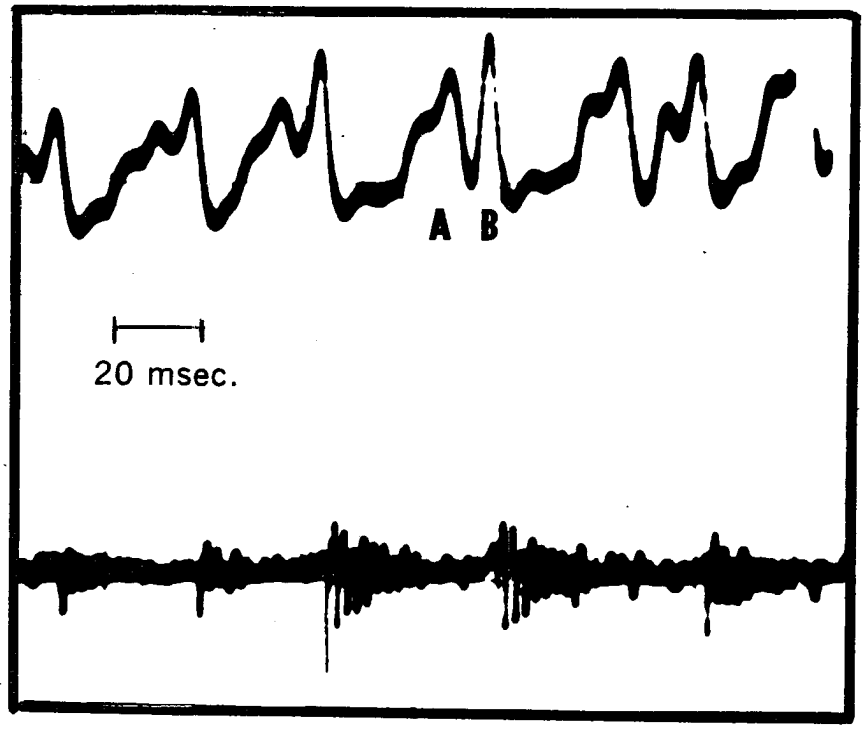


Figure 4. Glottogram (top) and microphone signal of steady-state "creaky voice" (also known as "glottal fry" or "vocal fry").

octave lower than the frequency of the tone, he concluded that "...le glottogramme est deux fois plus lent en période que la note émise." It is a problem for psychoacoustics to explain why subjects are sometimes off by one octave in attempting to match the pitch of their voice to the pitch of a pure tone, but acoustically the frequency of the vocal cord vibrations and the fundamental frequency of the sound emitted from the mouth are identical. In Fabre (1958) this misconception is not repeated.

Fabre's glottography differs from photo-electric glottography in that it measures vocal cord movements by measuring the variation in the transglottal impedance offered a weak, high-frequency electric current. But a brief discussion of his findings is pertinent here because his results are completely at odds with our observations on three points. We have found--along with those writers referred to above--that:

1. In normal voice, during sustained phonation, the closing phase of the glottal pulse is typically shorter and more abrupt than the opening phase.
2. Maxima in the speech sound wave correspond in time to minima, i.e., closed phases, in the glottal area curve. (There is a causal relationship between the first and second points.)
3. With increasing intensity of voice, the ratio of closed period to open period becomes larger. (cf. Sonesson, 1960, etc.)

Fabre (1957, 1958) finds the complete opposite:

1. The closing phase is typically longer and more gradual than the opening phase.
2. The maxima in the speech sound waves correspond to opening phases of the glottis.
3. With increasing intensity of voice, the ratio of the closed period to the open period becomes smaller.

There are several possible explanations for these serious differences with our findings, among them differences in the subjects used, or artifacts in the signal from the instruments used. However, several anomalies in the glottograms printed in these two articles offer strong evidence that Fabre simply misread and thus misinterpreted his data.

A casual inspection of the illustrations published in Fabre (1958) will show conclusively that all the oscillographic displays (glottograms plus simultaneous speech sound wave) in his Figure 1 were printed with the time dimension reversed (the sound waves decay from right to left as they are printed). With the resulting re-interpretation of these six glottograms, point #1 of Fabre (closing phase slower than opening phase) is thus reconciled to our findings (opening phase slower than closing phase). These glottograms do not offer the evidence Fabre claims they do for his point #2, and there is no variation in intensity represented so point #3 is not involved.

In his Figure 2 in the same article, which is identical to Figure 1 in the 1957 article, the time dimension is correct, but Fabre misread the open vs. closed glottis polarity, i.e., what he labeled "open glottis" is really closed glottis and vice versa. With the resulting re-interpretation of these glottograms they agree with our points #1, 2, and 3.

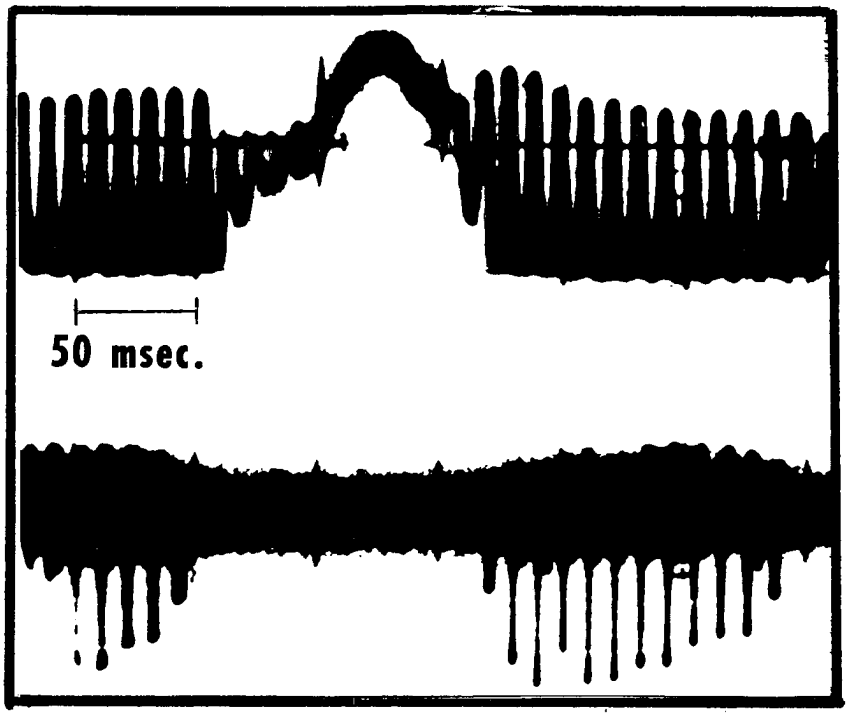
Figure 3 is more difficult to straighten out, but it appears that both the time dimension and the open vs. closed glottis polarity are reversed, with the exception of the left center glottogram, which appears to have the correct polarity in both dimensions, and which, therefore, offers visual evidence which is counter to Fabre's points #1 and 2. The other five, furthermore, offer no obvious evidence for his point #1.

A side-by-side comparison of these glottograms with those printed in Fabre and Frei (1959), which have the correct polarity in both dimensions, will quickly reveal the proper way to read the glottograms.

This sort of misinterpretation is not surprising given Fabre's apparently limited knowledge of acoustic phonetics. Fabre (1958, p. 773) states: "...sur les courbes [of the speech sound wave] en dehors d'une analyse harmonique instrumentale, l'oeil ne distingue pas les fréquences de résonance bucco-pharyngees mêlées a toutes les autres vibrations fournies par l'impulsion glottique." In fact a visual inspection of the speech sound wave in the top left corner of his Figure 3 is sufficient for us to estimate that the first formant is roughly 400 Hz: much too low for it to be the vowel 'A' [Fabre's symbol] as it is labeled, but is more appropriate for the vowel 'E'. The top four speech sound waves in his Figure 3 are thus mislabeled as to what vowel they represent. Furthermore, the article by Fabre and Frei (1959) abounds with misinformation on acoustic phonetics.

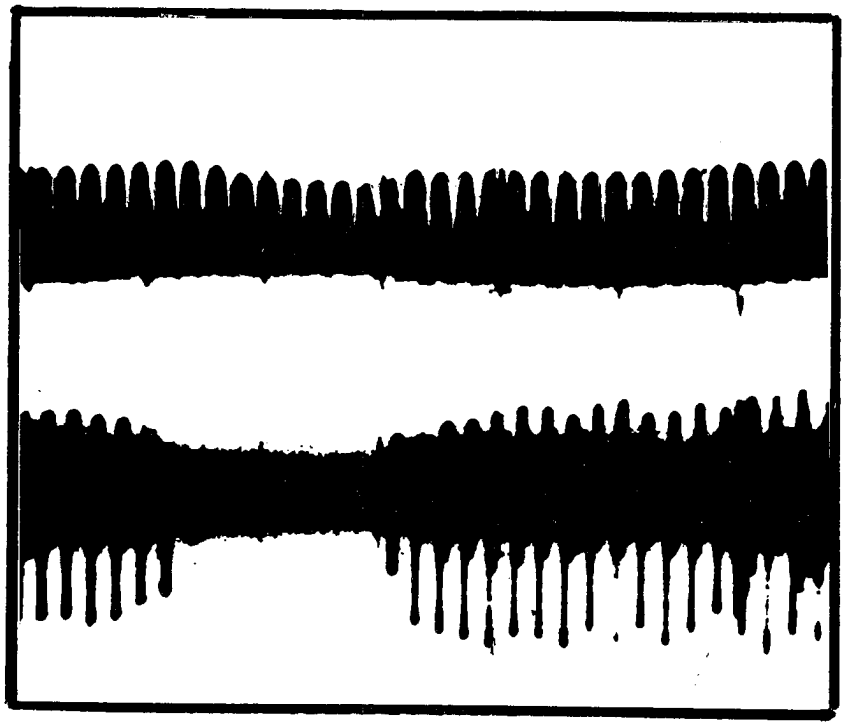
Returning to the discussion of our glottograms, Figure 4 shows a glottogram (top) and simultaneous microphone signal of steady-state creaky voice (also known as "glottal fry" or "vocal fry"). The amplitude of the glottal openings here are roughly 1/10th of those for normal voice, and the time scale here is much slower than before. Von Leden and Moore (1958), using high-speech motion pictures of the vocal cords, found the vibratory pattern of "glottal fry" to be characterized by a series of double glottal openings, separated from the next double opening by a relatively long closed period. The second of the two openings typically displayed a shorter duration, wider opening and a more abrupt closing phase. Although the irregularities in the glottogram in Figure 4 make any convincing interpretation difficult, it seems likely that the glottal openings marked A and B represent just such a double opening as von Leden and Moore reported. The glottal opening B has a shorter duration, wider opening and a more abrupt closing phase. Interestingly, only the second glottal opening at the moment of closure produces an acoustic pulse. The first glottal opening produces no sound because its closing phase is too gradual and thus cannot excite the vocal tract. If this interpretation is correct it explains why Wendahl, Moore and Hollien (1963) in their investigation of the acoustic and perceptual correlates of "vocal fry", found no evidence that double pulses were essential characteristics of "vocal fry". The double glottal openings do indeed exist in creaky voice, it seems, but the first of the pair of openings need not have an acoustic correlate. We often observed three glottal openings in quick succession, but again, only the last one at the moment of closure produced a sound pulse.

In Figure 5 is shown a glottogram (with greatly reduced time scale) of the sequence /a'pa/ abstracted from the larger sequence



(Retouched) Glottogram (top) and microphone signal of [ɑ'pɑ].

Figure 5



(Retouched) Glottogram (top) and microphone signal of [ɑ'bɑ]; same time scale as Figure 5.

Figure 6

/'pa'pa'pa'pa.../. For a /p/ in this environment it is clear that during the voiceless period, i.e., the period of non-vibration, the vocal cords move apart such that the maximum glottal area is greater than the maximum glottal area during vibration. The absolute size of this maximum area seems to vary as a function of the tempo of speech, being larger for a slow rate of speaking. In the transition from vibrating state to non-vibration, and vice versa, there are some vibrations of the vocal cords which are not picked up by the microphone. This could be due to the acoustic impedence of the lips which would drastically reduce the amplitude of the sounds arriving at a microphone in front of the mouth, or because these vibrations are too gradual and do not cause a pressure change large enough and rapid enough to excite the vocal tract (cf. Lisker & Abramson 1964).

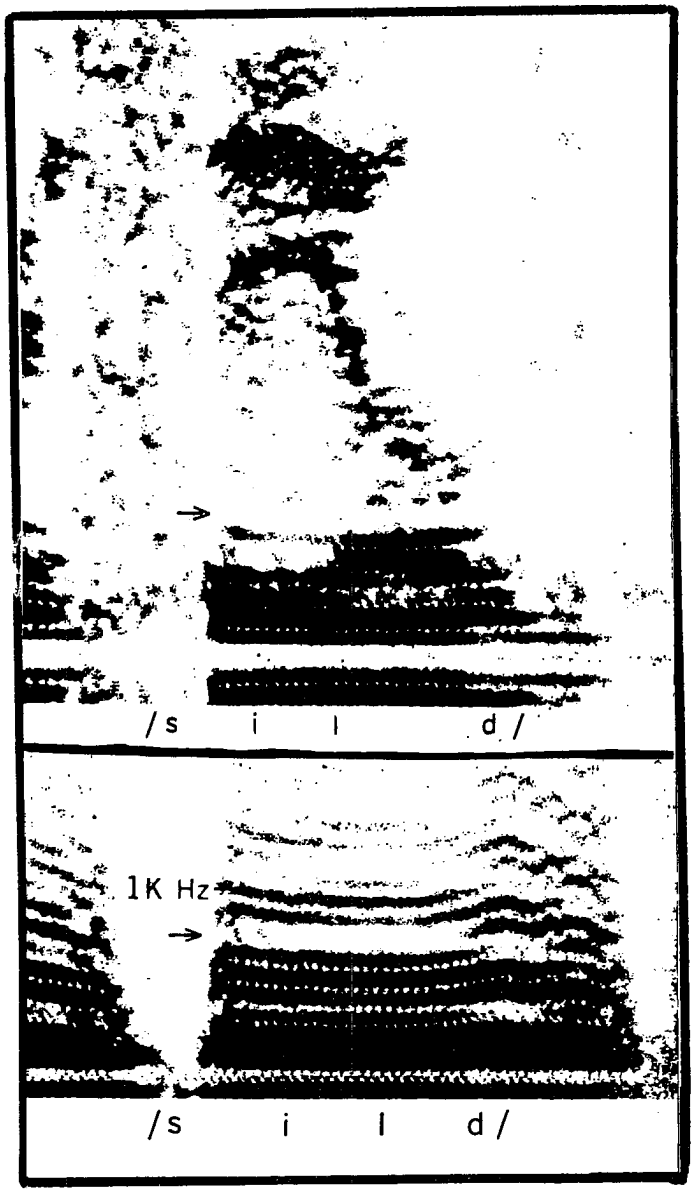
Figure 6 presents a glottogram for the sequence /a'ba/ taken from the larger sequence /'ba'ba'ba.../. In this case the vibrations continue throughout the duration of the closure, with a slight decrease in their amplitude due to the reduction in the pressure differential across the glottis.

Figures 5 and 6 are not necessarily offered as evidence on the state of the glottis during the English stops /p/ and /b/, but are merely illustrations of the type of the measurements that can be made with the glottograph. A systematic study of the stops of English and Korean is now under way.

Figure 7 is a spectrogram of the glottal area function. It is basic to current acoustic phonetic theory that the shape of the glottal area wave is almost identical to the shape of the glottal volume velocity wave (Flanagan 1962, 1965; Fant 1958). Thus we can treat a glottographic signal as if it represented the original sound produced by the glottis, unaffected by the resonances of the supra-glottal cavities.

Zeroes in the spectrum of the glottal wave (cf. Figure 7) have been noted by Miller (1959) and can be accounted for mathematically by Fourier analysis of the glottal wave (ibid; Flanagan 1962, 1965). Although there is still great uncertainty as to how these zeroes affect the perception of speech sounds and to what extent they contribute to identifying the speaker, their acoustic effect on the speech output is quite obvious as is shown by a comparison of the spectrograms of the glottal source and the speech output for the word "sealed" in Figure 7. The zero that occurs at the 7th harmonic (700 Hz) in the glottal source during the vowel is also manifested in the speech output at the mouth (see arrows).

Current acoustic phonetic theory is fairly well agreed that any acoustic coupling between the supraglottal cavities and the glottis is small and negligible. But our investigations did suggest that there is a very important mechanical coupling between the glottis and the supraglottal cavities, manifested every time the tongue or jaw is moved and every time a closure or release of a closure in the supraglottal cavity causes a sudden change in the pressure drop across the glottis. This mechanical coupling is responsible not only for changes in the frequency of vocal cord vibration but also for very real dynamic changes in the shape of the glottal wave and thus in the glottal spectrum. This can be seen in Figure 7 in that part of the spectrogram corresponding



Spectrogram of the word "sealed".

Simultaneous spectrogram of the glottal wave.

ng

Figure 7

to the closure for /d/, in which the zero previously located at 700 Hz now shifts downward to 620 Hz and a new zero becomes apparent at 350 Hz, thus reflecting a change in the shape of the glottal wave.

Perhaps the most useful feature of the photo-electric glottograph is the simplicity of its construction and use. Compared with much of the electronic gadgetry currently in use in speech research, the glottograph is about as straightforward to use as a meter stick. And as the necessity of sampling several physiological parameters of speech simultaneously becomes ever more apparent, researchers in the field will have to pay more attention in the design of their experiments to the ease with which the data may be obtained, recorded and analyzed.

I gratefully acknowledge the encouragement and helpful suggestions of Peter Ladefoged, Hans von Leden and Norris McKinney, and the considerable technical assistance of Stan Hubler and Ralph Vanderslice.

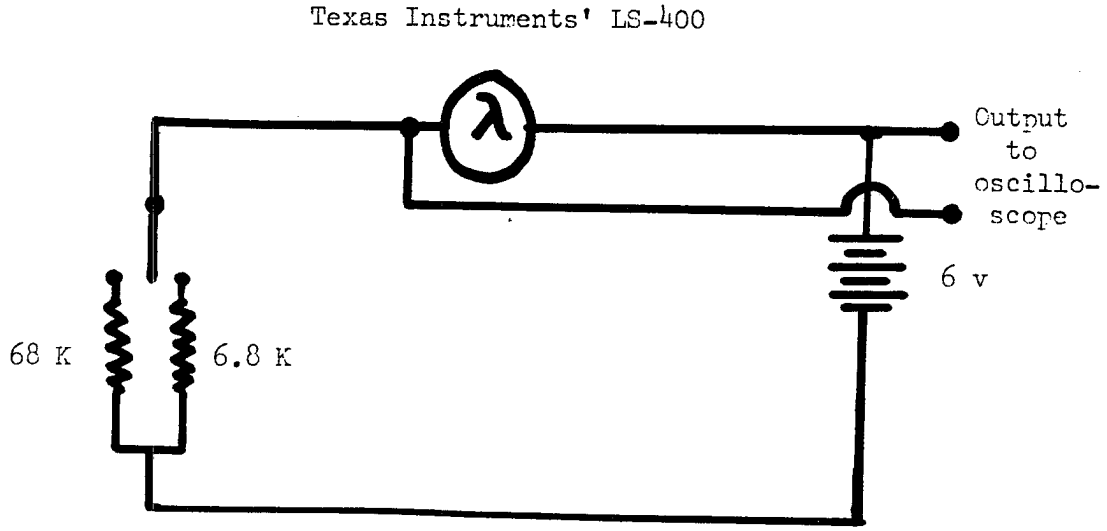


Figure A. Circuit diagram of glottograph.

Bibliography

Abramson, A. S., Lisker, L. & Cooper, F. S. (1965) "Laryngeal activity in stop consonants" Status Report on Speech Research-4 New York: Haskins Laboratories, 6.1-6.13.

Catford, J. C. (1964) "Phonation types: the classification of some laryngeal components of speech production" In Honour of Daniel Jones ed. by Abercrombie, D., et al., London: Longmans, 26-37.

Cederlund, C., Krokstad, A. & Kringlebotn, M. (1960) "Voice source studies" Speech Transmission Laboratory, Quarterly Progress & Status Report-1, 1-2.

Cooper, F. S. (1964) Discussions recorded in International Conference on Research Potentials in Voice Physiology, Syracuse, N.Y. [held in 1961] ed. by Brewer, D., New York: State Univ. of New York.

Fabre, P. (1957) "Un procede electrique percutane d'inscription de l'accolement glottique au cours de la phonation: glottographie de haute frequence" Bull. Acad. Nat. Med. 141, 66-9.

Fabre, P.(1958) "Etude comparée des glottogrammes et des phonogrammes de la voix humaine" Ann. Otolar. (Paris), 75, 767-75.

Fabre, P.(1959) "La glottographie electrique en haute fréquence, particularités de l'appareillage" Comptes Rendue, Société de Biologie 153, 1361-64.

Fabre, P. & Frei, A. (1959) "Analyse harmonique des glottogrammes et des phonogrammes de la voix chantee" Ann. Oto-Laryng. 76, 459-63.

Fant, G. (1958) Acoustic Theory of Speech Production Roy. Inst. Techn. Report 10, Stockholm.

Fant, G. & Sonesson, B. (1962) "Indirect studies of glottal cycles by synchronous inverse-filtering and photo-electrical glottography" STL-QPSR-4 1-3.

Flanagan, J. L. (1962) "Some influence of the glottal wave upon vowel quality" Proceedings of the 4th International Congress of Phonetic Sciences The Hague: Mouton & Co., 34-49.

Flanagan, J. L. (1965) Speech Analysis Synthesis and Perception, New York: Academic Press, Inc.

Holmes, J. N. (1962) "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter" paper delivered at the Fourth International Congress on Acoustics, Copenhagen.

Ladefoged, P. (1964) A Phonetic Study of West African Languages: An Auditory-Instrumental Survey Cambridge: Cambridge Univ. Press.

Lindqvist, J. (1965) "Studies of the voice source by means of inverse filtering" STL-QPSR-2 8-13.

Lisker, L. & Abramson, A. S. (1964) "A Cross-language study of voicing in initial stops: acoustical measurements" Word 20, 384-422.

Lisker, L. & Abramson, A. S., Cooper, F. S. & Schvey, M. H. (1966) "Transillumination of the larynx in running speech" abstract of paper delivered at 71st meeting of the Acoustical Soc. of Amer., Boston.

Miller, R. L. (1959) "Nature of the vocal cord wave" J. Acous. Soc. Amer. 31. 667-77.

Moore, P. (1962) "Observations on the physiology of hoarseness" Proceedings of the 4th International Congress of Phonetic Sciences The Hague: Mouton & Co., 92-5.

Moore, P. & von Leden, H. (1958) "Dynamic variation of the vibratory pattern in the normal larynx" Folia Phoniatrica 10, 205-38.

Sonesson, B. (1959) "A method for studying the vibratory movements of the vocal cords. A preliminary report" Journal of Laryngology 73, 732-37.

Sonesson, B. (1960) "On the anatomy and vibratory pattern of the human vocal folds -- with special reference to a photo-electric method for studying the vibratory movements" Acta-Oto-Laryngologica Supplement 156, Lund.

Wendahl, R. W., Moore, G. P. & Hollien, H. (1963) "Comments on vocal fry" Folia Phoniatrica 15, 251-55.

Zemlin, W. R. (1959) "A comparison of high speed cinematography and a transillumination-photoconductive method in the study of the glottis during voice production" unpublished M. S. Thesis, University of Minnesota.

UCLA Linguistic Phonetic Research, 1965-1966

Each student in the basic graduate phonetics course at UCLA (Linguistics 200) must write a term paper, incorporating at least a tape recording with text, concerning some point of interest in the phonetic structure of one or more languages. The Phonetics Laboratory keeps a file containing papers, recordings, and instrumental data produced by students in this and in other more advanced phonetics courses. We have found that this filing system is a great aid to students and faculty members engaged in research in phonetics. The files are cross-indexed, so that, for example, a given paper will be listed under at least the following headings: author's name, language investigated, phonetic problem studied, and type of instrumental data submitted with the paper.

The index is compiled strictly for our own use, and the quality of the work reported is, of course, quite varied. Nevertheless, we find that the index is helpful as a guide to the recordings and data filed in the laboratory. The index presented in the following pages is a continuation of that started in Working Papers in Phonetics, No. 2, covering the past academic year. The form of the entries is as follows:

Language — Phonetic problem

(Author of paper, course, date)

Language: Informant's name

Types of data on file

Summary

Index of Some of the UCLA Phonetic Research in 1965-1966

Timothy Smith

1. Amharic — Ejectives
(Language demonstration Ling. 200 March 1966)
Habte Mariam Marcos, Addis Ababa, Ethiopia
Tape with text

2. Amharic — Intonation
(Hudson, Grover Ling. 200 June 66)
Amharic: Teklu Neway
Pitch and waveform recordings, tape with text

Questions with question words have a rising/falling pitch contour, with highest pitch on the stressed syllable of the question word. Intonation questions have a rising intonation with the highest pitch on the stressed syllable of the verb.

3. Amharic — Labialization
(Tilem, Diana Ling. 200 June 66)
Amharic: Habte Mariam Marcos
Tape with text, spectrograms (8kc, WB)

Tentative conclusion is that there are two types of labialized consonants in Amharic: those with simultaneous lip rounding and those with sequential (post-consonantal) labialization.

4. Arabic — Emphatic consonants ṭ ḍ ṣ ṛ
(Barber, Lucie Ling. 200 Jan. 66)
Arabic (Lebanese): Nancy Sadka
Spectrograms (8kc, WB), oral pressure and waveform recordings, tape with text

Emphatic stops show greater oral pressure than non-emphatics, ṣ has lower freq. cut-off for noise than s (1500 Hz opposed to 2700 Hz), apparently no significant difference in vowel formant frequency after emphatic as opposed to non-emphatic Cs.

5. Arabic — Pharyngealized consonants
(Hatch, Evelyn Ling. 200 June 66)
Arabic (Sudan): Mr. Kheiri
Spectrograms (8kc, WB), amplitude and waveform recordings, tape with text.

Investigation of some of the acoustic correlates of pharyngealization in Sudanese Arabic.

6. Arabic — Vowels

(Anderson, Betty Ling. 200 Jan. 66)
Arabic (Syrian): Samir Habes
Tape with text

Some phonetic variants in the vowel system.

7. Bikol — Vowels
(de Castro, Rosenda Ling. 200 Jan. 66)
Bikol: Lilia Realubit
Photographs of lip positions for Vs, tape with text

Measurements of lip position (height and width of opening) for the Vs and an articulatory description of the vowels.

8. Burmese — Voiceless nasals
(Language demonstration Ling. 200 Feb. 66)
Maung Thein, Rangoon, Burma
Tape with text.

9. Castilian Spanish
See Spanish - Downey, Julie

10. Cebuano, Hiligaynon, Tagalog, Ilocano — Vowels
(Villanueva, Thelma Ling. 200 Jan. 66)
Cebuano: Fe Lucero
Hiligaynon: Author
Tagalog: Rosenda de Castro
Ilocano: Fe Medel
Spectrograms (8kc, WB), tape with text

Comparison of vowel length (in standard frames consisting of Spanish loan words) in the four languages.

11. Cuyonon — Vowels
(Timbancaya, Ester Ling. 200 Jan. 66)
Cuyonon: Author
Spectrograms (8kc, WB), tape with text

Acoustic and articulatory description of Cuyonon vowels.

12. Czech — d ɖ z ʃ
(Jensen, Janet Ling. 200 Jan. 66)
Czech: Kateřina Bednařova
Palatograms of above sounds, tape with text

Articulatory description of above sounds

13. Danish — Stød, vowels
(Language demonstration Ling. 200 April 66)

Ralph Brandi, Denmark

14. English, American — Vowels
(Reilly, M. Ling. 200 Jan. 66)
American English: M. Reilly, T. Smith, Karen Brown
Spectrograms (4kc WB), pitch recordings, tape with text

American English vowels in the frame /h_d/ were recorded under normal and helium atmosphere. First and second formants increased in frequency under helium, while maintaining relative position.

15. English, American — (Speech synthesis)
(Whitaker, H. Ling. 200 June 66)
American English: Author served as model for the utterance to be synthesized.
Spectrograms (8kc, WB; 4kc, WB & NB) of model, spectrograms of synthetic utterance, amplitude and waveform recordings, tape loop.

Utterance was recorded ("Please feel the leaves"), measured, and a sleeve for the UCLA synthesizer was drawn. The synthetic utterance was readily intelligible. Some procedures for sleeve-drawing are given in the paper.

16. English — Intonation
(Londe, D. Ling. 200 Jan. 66)
Pitch and amplitude recordings

A comparison of three recorded renderings of Hamlet I.2.

17. English, American
See Mandarin - Brown, Karen

18. English, Spanish, French — r-type sounds
(Powers, Dolores Ling. 200 Jan. 66)
American English: G. Powers
British English: M.A.A. Tatham
Spanish: R. Cabrera
French: R. Parent
Spectrograms (8kc, WB), tape with text

Acoustic and articulatory study of r-type sounds in these languages.

19. French
See English - D. Powers

20. French — liason
(Cabrera, R. Ling. 200 Jan. 66)
French: Suzanne Lemaire (Montreal, Quebec)
Tape with text

Liason in Canadian French.

21. German — stops (word-final)
(Menzel, P. Ling. 200 Jan. 66)
German: from Zurich, Vienna and Frankfurt, not named
Spectrograms (8kc, WB), tape with text

Tapes were spliced and played back to informants. Word-final stops usually devoice, although informants could distinguish voiced and voiceless (phonemically) stops in word-final position. A tense/lax distinction supposedly distinguishes word-final stops.

22. Gujerati — Voiced aspirated stops, murmured vowels
(Language demonstration Ling. 200 March 66)
P. Patel, Baroda, Gujerati, India
Tape is available.

23. Gujerati — Murmured vowels
(Anderson, Geraldine Ling. 200 Jan. 66)
Gujerati: P.J. Mistry
Spectrograms (8kc, WB & 4kc, NB), tape with text

Murmured and voiced vowels contrast in Gujerati. Formants not so clearly defined in murmured vowels.

24. Gujerati — Breathy (murmured and non-murmured vowels in normal and whispered speech
(Patel, P.G. Ling. 200 June 66)
Gujerati: Author (native speaker)
Spectrograms (8kc, WB), tape with text

25. Hiligaynon
See Villanueva, Thelma - Cebuano

26. Ilocano
See Villanueva, Thelma - Cebuano

27. Japanese, Ryukyuan
(Inamine, S. Ling. 200 Jan. 66)
Japanese & Ryukyuan: Author
Tape with text

Some phonetic differences between the two.

28. Ladino
See Downey, Julie - Spanish

29. Mandarin — Whispered speech
(Morton, Palmyra Ling. 200 Jan. 66)
Mandarin: Liang; Chen
Spectrograms (8kc, WB & NB), tape with text

30. Mandarin, English r
(Brown, Karen Ling. 200 Jan. 66)
Mandarin: Jeffrey Cheung: Peter Chang
English: Grover Hudson
Lip position photographs, electromyographic data, tape with text

English /r/ /__V shows more action potential (obicularis oris) and more lip rounding than Chinese /r/ same position.

31. Mandarin Intonation
(Jones, Josette Y. Ling. 200 June 66)
Mandarin: Author (native speaker)
Tape with text, pitch recordings (Oscillomink & pitch meter) and oscillograms.

Questions in Mandarin have a slightly higher pitch level than the equivalent statements, and where no question words are present, the last syllable of the question is of longer duration.

32. Mandarin Pitch
(Landerman, P. Ling. 200 June 66)
Mandarin: Josette Y. Jones
Pitch, oral pressure and waveform recordings

Recordings of [paʔ], [paw], and [pa] with various tones. Oral pressure before release of [p] is taken as indicative of subglottal pressure before stop release and during vowel articulation. In general there is a direct relation between the two phenomena, subglottal pressure and tone.

33. Quechua Stops
(Slayton, J. Ling. 200 June 66)
Quechua: Francisco Tapia, Danial Tapia
Pitch, waveform, and amplitude recordings, tape with text

34. Ryukyuan
See Inamine, S. - Japanese

35. Spanish
See Powers, Dolores - English

36. Spanish Consonants
(Fierro, G. Ling. 200 Jan. 66)
Spanish (Ecuador): Mrs. F. Fierro
Tape with text

Phonetic description of the consonants of this dialect.

37. Spanish, Ladino
(Downey, Julie Ling. 200 Jan. 66)

Spanish & Ladino: not named
Tape with text

Comparison of the two languages.

38. Swedish
(Dillingham, D. Ling. 200 Jan. 66)
Stockholm Swedish: Kerstin Sandstrom
Norrland Swedish: Britt Sheikholeslami
Spectrograms (8kc, WB), tape with text, pitch recordings

Comparison of the two dialects.

39. Tagalog iw ew aw ow
(Cabrera, Neonetta Ling. 200 Jan. 66)
Tagalog: Author
Spectrograms (8kc, WB), lip position photographs, tape with text

Comparison of the diphthongs in terms of lip position (height and width of opening) for both the vowel nucleus and the glide.

40. Tagalog
See Villanueva, Thelma - Cebuano

41. Taiwanese
(Wang, Shirley Ling. 200 Jan. 66)
Taiwanese: C.S. Wang; K.R. Chuang
Spectrograms (8kc, WB), tape with text

Mainly articulatory description of some of the phonetic features of the language.

42. Turkish Stress and Pitch
(Freeman, Nancy Ling. 200 June 66)
Turkish: Ibrahim Karal
Pitch and amplitude recordings (and waveform), tape with text

43. Vietnamese
(Thang, T. T. Ling. 200 Jan. 66)
Hanoi Viet.: Author; Saigon Viet.: Nguyen Dang Long
Tape with text

Description of some of the differences between the dialects.

44. Yoruba Nasal and oral vowels
(Banjo, A. Ling. 200 Jan. 66)
Yoruba: Author
Spectrograms (8kc, WB), tape with text

Treats the distinction between nasal and oral vowels in Yoruba.

45. Zulu — Clicks and ejectives
(Language demonstration Ling. 200 Feb. 66)
Anthony Ngubo, South Africa
Tape with text

Notes on Linguistic Fieldwork

We are examining the phonetic structure of a wide variety of languages. Much of this work arises from research on specific phonological problems. But we are also interested in developing an adequate general phonetic theory which will enable us to give valid descriptions of the phonological contrasts which occur in the languages of the world. Collecting data for this theory often involves field trips, so that we can record the speech of informants who cannot be brought into the laboratory. There is no difficulty in making high quality tape recordings which may be later analysed with instruments such as the sound spectrograph. But the techniques for recording physiological data in the field have not been so fully developed.

One of the most useful ways of recording data concerning the positions of the articulators is by means of palatography. The system that we use is basically that designed at the University of Edinburgh (Anthony 1954). A dark powder is sprayed on the upper surfaces of the informant's mouth, and he is then asked to say a single word in which there is only one consonant contact between the tongue and the upper teeth or palate. When the tongue touches the roof of the mouth it wipes off some of the powder. The contact areas can be studied and photographed with the aid of a system of lights and mirrors; the resultant record is known as a palatogram. If the subject sticks his tongue out immediately after saying the word which is being examined it is also possible to see (and photograph) the parts of the tongue which have touched the powder on the roof of the mouth. This kind of record is called a linguagram. The tongue does not have the same shape when it is protruded as it does during the articulation, so linguagrams cannot be interpreted too precisely. But, as has been shown in Ladefoged (1957), palatograms of articulations in the front of the mouth can be interpreted with a fair degree of precision as long as there is a dental impression of the roof of the mouth available, so that the contours and sagittal section are accurately known.

For the convenience of linguists working in the field we have constructed a portable palatography apparatus which folds up so that it can be carried in a case only a little larger than the camera and flash unit. (Ralph Vanderslice designed and built this device.)

Another useful kind of physiological data is a record of some aspects of the airstream dynamics, such as the pressure of the air in the mouth or below the vocal cords, and the rate of flow of air out of the mouth. It has been shown (Ladefoged 1964) that these records provide a great deal of information about the formation of sounds such as the double stops /kp, gb/ and the implosives which occur in African languages and elsewhere.

In order to record these data in the field we have devised a system for recording variations in air pressure on tape along with the audio signal. (Stanley Hubler designed and built this device.) The pressure is transduced with the aid of a moveable anode pressure transducer and the resulting voltage variations converted into frequency variations usually in the range from 11 to 15 kHz. This signal is recorded along with the regular audio signal on a Nagra portable tape recorder which tests have shown to have a frequency response of 40-20,000 Hz $\mp$ 1.5 db.

The entire system is operated from the rechargeable batteries of the tape recorder. On returning to the laboratory the pressure signal can be demodulated and displayed on an oscilloscope or written with the audio waveform on an ink-writing oscilloscope such as an oscillomink. Alternatively, at the time of the original recording, the pressure may be superimposed on the audio signal in the range between 5 and 8 k/cs. This record may be analysed, without further processing with a sound spectrograph. Since the pressure varies in an interesting way at precisely the times when there is no audio signal, the presence of the two signals on the one spectrogram is not too disturbing.

P.L.

Bibliography

Anthony, J. (1954) "A new method of investigating the tongue positions of consonants" Science Technologists Bulletin 2-5.

Ladefoged, P. (1957) "Use of palatography" J. Speech and Hearing Disorders 22.5, 764-74.

Ladefoged, P. (1964) A Phonetic Study of West African Languages Cambridge University Press.