ACOUSTICS OF SPEECH*

Osamu Fujimura

* Precongress paper submitted in February 1971 as material for discussion at the Symposium on Speech Production-Speech Perception: Their Relationship to Cortical Functioning, Vancouver, April 13-15, 1972. This paper will be published in the proceedings of the symposium, edited by John H. Gilbert. The author wishes to acknowledge with deep gratitude that numerous corrections and advices on the manuscript have been contributed by Arthur S. House, who served as a discussant at the session, and the exposition has been improved accordingly for this revised version.

Introduction

The problems I shall discuss in this paper are those of speech phenomena in more or less direct relation to linguistic units and their functions. Ultimately, our concern will be to understand the essentials of the relation between the linguistic codes and the acoustical (and related) properties of speech phenomena, but for this to be achieved, it is first necessary to outline acoustical theory and facts that restrict the nature of the phenomena we will observe. I will refer to some applicational problems, only as criteria of our understanding and hypotheses. For obvious reasons, I will have to restrict myself to reorganizing the vast problem showing crucial points by concrete examples, without going completely into individual details, which are also necessary for actual speech studies.

As is well known, speech phenomena are manifestations of linguistic codes. It must be emphasized, however, that not all the information conveyed by speech waves is to be related to the linguistic entity. We can judge, for example, some physical characterizations of the speaker, whether male or female, old or young, and probably even tall or short to some extent, by merely listening to his voice, even when the utterance does not involve much semantic or syntactic idiosyncracy. The question of where to place the boundary between linguistic and non-linguistic aspects of speech being put aside for the moment, it is perfectly clear that not all the information of speech waves is to be treated in any structural description of language. Still, for us to discuss the characteristics of speech, it is important to know about these extra-linguistic factors to some extent, even when our concern is only to understand the acoustic correlates of linguistic (phonological) structural units. This is so because what we observe as natural phenomena inevitably involves those extra-linguistic factors, and therefore somehow we will have to be prepared to exclude the effects of these irrelevant factors by some sort of abstraction, or filtering, let us say. This sort of filtering is nothing at all like any mechanical process commonly known as electronic techniques, but something quite intricate, but nevertheless quite commonplace in the daily mental processes of a human being (even though it unfailingly escapes his

own appreciation).

Looking at speech phenomena carefully from an observational point of view, we find first of all that they are variable or "noisy" in a broad sense of the word. Variability or fluctuation, as a matter of fact, seems to be one of the very essential characteristics of human performance, particularly in contrast with what machines do. The signal to be received being always noisy, a human being is somehow equipped with an effective means of extracting the signal from noise, and also of uncovering the code structure of language, notably, from the limited noisy experience. This is a remarkable, and indeed intriguing fact not only for the engineer but also for the speech scientist and the linguist because it appears that there is no algorithm to be given for duplicating this ordinary human ability. Even though I agree completely with the Chomskyan thesis that a rigorous description of the linguistic regularity has to be based on a totally deductive methodology, it is at the same time true that, from a heuristic point of view, our science needs faithful recording of natural phenomena and inductive data processing as much as it is practical. Since we have not yet established, or, as a matter of fact, even proposed the entire chain of descriptive levels from syntactic codes down to acoustic waves, we need to narrow down the domain of search by whatever is known.

One, and perhaps the only, way to cope with the complexity of our problem in an empirical approach, before having even a tentatively proposed overall outline of structural description, is to try to find a set of observational samples that correspond to different values of a certain dimension under control, and to keep all availables in other dimensions in the "same" condition as parameters, even though we do not know exactly what these may be. This method of minimal contrast (or comparison ceteris paribus as Roman Jakobson puts it), is probably the most practical linguistic advice which one must follow in designing experiments in speech science. In order to do this, however, we still need to have some rough idea at least, about what kinds of control dimensions might come into the picture. In some cases we need to know exactly about certain relevant (e. g. phonological or phonetic) constraints on combinations of the controlled variable values in different dimensions; if not we may try to let the

subject produce, or identify, phonologically impossible or peculiar forms, for example, and he may not be performing in his normal mode of language use.

One particular case of "noise" in the broad sense outlined above is the random variability of human performance in a given linguistically well-defined task. In production of a certain sentence, for example, it is not surprising at all that a speaker does not repeat the same physical utterance twice.* This random variability of the same linguistic material by the same talker, for example, may be eliminated from observation only through some kind of statistical processing. In order to make this statistical process meaningful, however, we need to set some, not necessarily obvious, framework for describing the signal. This often has to be based on theoretical considerations of the physical production mechanism because we have no other way of understanding the nature of the signal.

By way of illustrating the point at issue, let us consider an extremely simple case--that of statistically processing different utterances of a monosyllabic word in isolation, which may exist in some language and be transcribed as [pu]. Suppose we have an impulse-like waveform followed by a short interval of silence and then a quasi-periodic [u]-sounding waveform. After making sure by listening that we have a sufficiently convincing sample of [pu], let us move the position of the first impulse-like waveform so as to vary the duration of the silent interval. Within a certain range of duration values, we will continue to hear [pu]'s. Let us make a sufficiently large number of such [pu]-sounding waveforms that have a normal distribution around a representative value of the silence duration, and simply add them up as time functions. It will be easy for acoustic phoneticians to believe that the resulting waveform, which is an average of the many [pu]'s, will sound like a good sample of [ɸu] or [fu].

This gedankenexperiment can be reinterpreted easily as a case of analyzing natural speech samples spoken as [pu]. If we did not know the

---

* The auditory image of a listener to the same stimulus, likewise, will not be the same, even though the different but redundant signals certainly will lead to the perceptual identification of the same linguistic code.

relevant variable for discussing inherent characteristics of the consonant, and if we blindly dared to take the statistical average of the observed waveforms for a set of utterances of [pu], obviously we would obtain the wrong conclusion. What we should do instead in this case is to measure the relative timings of the occurrence of the impulse-like wave relative to the initiation of the vocalic waveform, and take the average for this parameter value separately from others that may be processed similarly. Then a reconstruction of the "average" waveform will surely sound like a representative [pu].

How can we know, then, if we are doing the right thing in each case of exploring the relevant characteristics of speech sounds? Are we simply trying to give a meaningless answer to a circular problem? It is quite clear that truly inductive methodology cannot survive if we wish to be logically rigorous in our research principle. Even seemingly inductive experiments have to be guided by theoretically insights. There was good reason for scientists in general speech studies to depend heavily on an acoustic theory of speech production and speech synthesis experiments, in analysis and interpretation of speech waves. Of course it is true that the invention of the sound spectrographic technique for speech analysis led us to a new epoch of speech research at the beginning of the latter half of this century, but still it had to be supplemented by these deductive methods of investigation, in order to achieve our present understanding of speech phenomena.

1. Acoustic Waves

The speech signal, in the physical form of a sound pressure wave, carries the verbal message from the speaker to the listener. The speaker also monitors the actualization of his message through this signal as the final product, as well as through some other physiological feedback channels. It is important to recognize, particularly for the engineer designing a system for processing speech signals, that this form of signal as the output from the speaker is never the same as the acoustic signal that the listener, or any receiving device, receives at a far distance in a room, as in most real conversations. The acoustic distortion introduced by this process

of transmission is not insignificant, and actually no known device, however sophisticated, can restore the original signal exactly, even though the distortion is certainly that of a strictly linear process. It is amazing that a human listener can communicate in ordinary rooms even with a fair amount of noise. This fact is by itself strong enough proof that the speech signal as it comes out of one's speech organs is highly restricted in physical characteristics, viz. that it is highly redundant as an acoustic signal, and also that the human process of verbal perception is not something of a simple nature that one can lightly attempt to simulate by a machine (particularly if general performance for a wide range of message material is sought). I must emphasize that when I talk about the nature of speech waves I am speaking about an ideal situation, where, for example, the sound signal is picked up by a high-quality microphone in an anechoic chamber. The characteristics of the perceptual process as mentioned above are of an inherently different nature compared with the physical phenomena we will talk about here. Today we can demonstrate some machines that, for example, can recognize spoken digits almost perfectly even in a noisy laboratory environment, provided that the speaker's own production of the material to be recognized is processed and stored beforehand. This kind of achievement is certainly not to be expected from a straightforward engineering scheme that pays no attention to the essential characteristics of speech.

1. 1. Acoustical Theory of Speech Production

An acoustical theory of the speech production process was developed first by M. Kajiyama, a physicist, based on his radiographic experiments conducted in cooperation with T. Chiba, a linguist. The results were reported in a book which was published in English about thirty years ago (Chiba and Kajiyama 1941). He applied Webster's horn equation to the acoustic system, the vocal tract, which is formed by the speech organs extending from the glottis to the lips. Based on midsagittal dimensions obtained from laterally taken x-ray photographs and some other necessary estimates of the cross-sectional shapes of the tract, he derived, through approximate calculations, the formant frequencies for the five Japanese

vowels. In order to corroborate the validity of the calculations, the vowel sounds were also acoustically synthesized for listening evaluations.

Independently of this early work, workers at the MIT Acoustics Laboratory, including Gunnar Fant from Sweden, took up the same problem and proposed electrical circuits simulating the acoustic system (Stevens et al. 1953). Based partly on this simulation experiment and partly on numerical calculations of the acoustic tube, by the same physical principle as in Kajiyama's work but assisted largely with network theoretical considerations, Fant completed a substantial and well developed work comprising a new approximation theory for vowels and theoretical and experimental materials for different kinds of consonants. This work of his was published in 1960 as the first comprehensive treatise on speech acoustics (Fant 1960).

The starting point in the modern theory of speech production is to regard the vocal tract as a linear acoustic transfer system (a four-terminal). It is assumed that there is no coupling between the source generator and the transfer system, an assumption which is not really true but works quite well in most cases that concern us. The additional assumptions of plane wave propagation, rigid walls and a smooth monotonic spectrum envelope for the voice source are not exactly true either, from a rigorous point of view, but all work well for many purposes. The merit of these assumptions has been established through various sorts of experimentation, and we wonder what would have happened if we did not adopt them to start with. We have learnt in many fields of physics that a grossly approximate, but at the same time comprehensible, basic framework is indispensable for data interpretations and discussions that lead to the progress to science.

1. 1. 1. Spectral and Temporal Aspects

The acoustic speech signal is characterized in two aspects, viz. in time and frequency domains. In terms of their acoustic correlates individual vowels are more strongly characterized in the frequency domain by their inherent spectral properties, whereas consonants generally cannot be characterized without specification of their temporal patterns. As is well known, the sound-spectrographic technique of speech signal analysis has made it possible to represent and record temporally changing spectral

patterns for visual inspections of the signal characteristics. There is an inherent and general sort of problem about simultaneous specification of both temporal and spectral structures of a given signal, and one of the salient successes in the design of the spectrograph for speech analysis was the adoption of a rather loose frequency resolution, viz. 300 Hz for the filter bandwidth (Potter et al. 1947).

In treating the general problem mathematically, it is convenient and comprehensible to represent the characteristics of the signal and those of the system by Laplace transforms, with time serving as a relatively slowly changing parameter after the transformation. By doing so, the output signal from the articulatory system can be represented simply as a numerical product of two functions representing the source signal and the transfer function of the vocal tract, respectively. If the vocal tract is excited solely by the vocal cord vibration, as in the case of vowels and some other voiced sounds, we can consider the volume velocity through the glottis as the source signal. In such a quasi-static representation, therefore, we can relate the sound pressure picked up at a certain distance from the mouth p(s) to the product of three terms: the volume velocity at the glottis $u_g(s)$, the transfer function of the vocal tract T(s) defined as the ratio of the output volume velocity to the input volume velocity, and the radiation characteristics represented by R(s), as follows (s represents frequency in the complex number domain):

$$p(s) = R(s) \cdot T(s) \cdot u_g(s) \qquad (1)$$

Among the three terms, T(s) is the determinant factor for phonetic values.

1. 1. 2. The Vocal Tract

In the case of vowels, we assume that the vocal tract is an acoustic tube with no branching and with the sole excitation applied at its closed glottal end. According to Fant's theory (Fant 1960, 1956, 1968), the transfer function T(s) can be decomposed into separate formant terms $F_i(s)$, each representing a simple-tuned resonance:

$$T(s) = \prod_{i=1}^{n} F_i(s) \qquad (2)$$

where n is supposed to be sufficiently high, or otherwise the expansion can be truncated at a certain point beyond which a so-called higher-pole correction term represents the effects of the remaining terms. Each term $F_i(s)$ for the i-th formant can be represented as

$$F_i(s) = \frac{s_i \cdot s_i^*}{(s - s_i)(s - s_i^*)} \qquad (3)$$

where s_i^* denotes the complex conjugate of s_i, and s_i represents the formant frequency and the damping factor, in general. This s_i (and s_i^*) represents a "pole" of the transfer function in circuit theory terminology.

In the case of consonants, inherently transient characteristics have to be treated as temporally changing phenomena, of course, and the above treatment of vowels provides the basis of the notion of formant transition. Namely, the set of formant frequencies, which, after Fant, may be called the F-pattern including, let us say, up to the third formant (in consideration of its direct perceptual effect), changes in time in a manner that is characteristic of the sequence of phonetic segments.

For a more-or-less stationary time segment manifesting a consonant in the acoustic signal, we can generalize the expression of T(s) as:

$$T(s) = \prod_i F_i(s) \cdot \prod_j A_j(s), \qquad (4)$$

where $A_j(s)$ has an inverse characteristic of $F_i(s)$, viz.

$$A_j(s) = (s - s_j)(s - s_j^*) / s_j \cdot s_j^* \qquad (5)$$

This term, characterized by the zero s_j of the transfer function of the vocal tract, may be called the antiformant (Hattori et al. 1958). An antiformant can be observed as a valley in the speech spectrum and represents a spectrographic consequence of a consonantal feature that is characterized by a particular way of deviating from vowels in terms of some topological conditions of the acoustic system, i. e. either branching of the vocal tract or location of the excitation source somewhere midway along the tract. Usually the spectral valley itself does not lead to significant perceptual effects, but is important as a structural determinant of the entire spectral shape of the output sound. A vowel formant can be annihilated by an anti-

formant, in a case such as nasalization, giving a perceptual effect of deviation from the vowel-like quality (Fujimura and Lindqvist 1971). The introduction or annihilation of an antiformant as a result of a topological change (switching of relevant channels) in the acoustic system of speech production in cases such as [mi], and perhaps [li] also, can cause an apparent discontinuity in the spectrogram (Hattori et al. 1958).

The pole-zero locations of the vocal-tract transfer function can be calculated if the cross-sectional area of the vocal tract is given as a function of position along its longitudinal axis (the area function). It is difficult, however, to estimate the damping factors (real parts of poles and zeros) and the distribution and nature of the turbulent noise sources exactly with the physical data presently available.

1.2 Direct Measurement of the Transfer Characteristics

Fant's theory of the transfer function for vowels was proposed early in the 1950s, and its validity has been attested through synthesis experiments. The theory predicted that a series connection of simple-tuned resonance circuits, in accordance with the formulae above, would provide us with complete control of the phonetic value in the entire range of vowel qualities, and experiments have corroborated that natural-sounding vowels of clear phonetic values can be produced by controlling only the lowest three formant frequencies, all the rest of the parametric conditions being left constant.

More recently, acoustic measurements of the vocal-tract characteristics have provided another experimental support of the theory (Fujimura and Lindqvist 1971) The point of this study is that in observing the speech signal we always have data which reflect an inseparable combination of the source and transfer characteristics. One of the two factors at least must be substantiated independently from the other through direct observations. In our experiment, for this purpose, an artificial sound source, in a form of sweeptone, was applied transcutaneously to the vocal tract of a normal subject just above the closed glottis. Assuming that the transfer characteristic through the thin neck wall, which is unknown, did not change from one vowel articulation to another within a series of measurements in a short experimental session, we compared the frequency response curves of different

articulations quantitatively, and tried to see if, in conformity with the theory, the difference could be interpreted as due only to different F-patterns.

Even though in this method of study there is still an unknown factor that is included in the data we can measure, all we need to assume is that this unknown factor remains constant. This assumption if independent of a similar one for the voice source in the case of comparing speech waves for different articulations of the same subject. Also, in this case of sweeptone excitation, we obtain accurate frequency-response curves rather than only a set of frequency-sampled values, the latter being the case when we observe the sound spectrum with its harmonic structure due to periodic glottal excitation.

The actual procedure for quantitative interpretation of the data curves was as follows. Let us denote the sound pressure amplitude of the sinusoidal wave picked up in front of the subject's mouth orifice by D(f) as a function of frequency f. This function is assumed to be decomposed into the product of the formant functions F_i, with the formant frequency f_i and the formant bandwidth b_i as parameters, multiplied by a common fixed frequency function C(f), viz.

$$D(f) = C(f) \cdot \prod_i F_i (f; f_i, b_i). \qquad (6)$$

For each of the frequency response curves recorded for different articulations, the second function on the right (the product form) was realized by a series connection of simple-tuned circuits, their parameters f_i's and b_i's being controlled for the best match between both sides of (6), using an ad hoc but fixed function C(f). When a satisfactory match was obtained, the parameter values were recorded as the estimated formant data for the observed articulations. The function C(f) was determined by a successive approximation, starting from a first guess and adjusting it to minimize the average mismatch for all the data curves in a set of measurements for which it is supposed to remain fixed, in a course of repeated matching processes.

This process of interpretation as a deductive technique of analytic measurement is called "analysis by synthesis," and was discussed by

K. N. Stevens and Morris Halle (Halle and Stevens (1962, Stevens 1960) as a means of spectral analysis of speech sound that might serve as a model of human speech perception. In a series of experiments conducted at MIT, data were derived for spectral structures of vowels, nasal murmurs etc., by applying this method in conjunction with the use of an interactive computer (Bell et al. 1961, Fujimura 1962, Paul et al. 1964). The sweeptone measurements as described above represent an analogous study, and provide us with, for example, accurate estimates of formant bandwidth values for vowels; it was revealed, contrary to what had been believed earlier, that typical bandwidth values for the first formant are larger when the first formant frequency is low than when it is in the middle range. This finding constituted another experimental proof of the acoustical significance of the compliance of the vocal tract walls, which had been inferred by Fant and Sonesson from analyses of speech in high pressure environments (Fant and Sonesson 1964). At the same time, by empirically deriving the fixed function C(f) for the particular measurement conditions through vowel samples, we also estimated unknown frequency characteristics of the articulatory system for some (stationary consonantal gestures. Resonance frequencies of the vocal tract during complete closure conditions for different stop consonants were also derived, and the theory of nasalization and nasal murmurs (Hattori et al. 1958), Fujimura 1961c, 1962) was also supplemented with new quantitative data.

1. 3. The Sources

In the case of vowels, liquids, nasals and semi-vowels, the source of acoustic excitation of the vocal tract is primarily the glottal modulation of the air flow through the vibrating vocal cords. When there is a narrow constriction somewhere along the vocal tract as the air flows through, turbulence is produced and it serves as an excitation source. The location of this source varies according to the place of articulation, and the spectral shape of the output signal varies depending on this location even if the turbulence itself has the same spectrum. In general, if the vocal tract shape is kept the same, the formants are located at the same positions in the frequency domain regardless of the place of excitation, but locations of anti-

formants vary, and so does the resultant spectrum.

The turbulence noise as the excitation source, typically in the case of voiceless fricatives, is called frication (Fant 1960). An explosion of a stop consonant quite often is accompanied by a short frication, immediately after the impulse-like transient waveform introduced to the measured sound pressure signal by a sudden release of the articulatory closure. When in the articulatory movement, the constriction is opened up to a certain cross-sectional area, for example, towards the following vowel after a voiceless stop, and if the glottis remains appreciably open without having the vocal cords set in vibration, a fair amount of direct-current air flow passes through the glottis, and some turbulence noise is produced there. This excites the acoustic system and produces an h-like sound, and this effect is called aspiration. The transition between frication and aspiration is not necessarily clearcut; it is presumably an automatic consequence of articulatory movement, as when, for example, a vowel articulation follows a voiceless aspirated stop. The laryngeal (and pulmonic) gestures determine the extent of aspiration, i. e. when the glottis is narrowed and voice onset takes place.

It has been said sometimes that [h] is characterized by an inherently narrow glottal constriction, thereby producing turbulence noise. According to recent findings by direct observations of the glottal images during consonantal articulations (see § 2. 2), this notion seems to be misleading. The glottis can be wider than that for [s], for example, in the same environment (Sawashima 1968), and it can be said that [h] is one of the sounds that require the widest open glottis, even though it still requires a constriction, compared to the respiratory condition, and this constriction may be narrower than any constriction of the vocal tract. Incidentally, in [h] the vocal cords tend to maintain vibration in spite of being distinctly abducted, presumably, in part at least, because of the higher rate of air flow (compared to [s], for example).

In Japanese, it has been described that the phoneme /h/ is manifested as [h] when followed by the vowel /e/, as [ç] when followed by /i/, and as [ɸ] when followed by /u/ (cf. Hattori 1951j). Even though this apparently is true phonetically, it is questionable whether these allophonic variations

should be interpreted as a consequence of context-sensitive phonological rules of Japanese. In other words, it seems possible that we can interpret these phenomena as consequences of physical rules assuming the same control gesture for /h/, except that the articulatory conditions are left unspecified by this segment but are conformed with the gestures for the following vowel. The main source of turbulence noise is located either at the place of the articulatory constriction or at the glottis, or perhaps along an extended region along the vocal tract, depending on which constriction is aerodynamically the most favorable for the turbulence to take place. If this view is correct, then the next question will be what happens in the case of English, for example, where /hi/ is often described as [hi]. Whether the wider articulatory constriction for the vowel should account for the difference between the two languages, and whether there should be different temporal characterizations either for the particular consonant or consonants in general, in terms of e. g. some time constant, has to be determined.

1. 3. 1. Vocal Cord Vibration

Recently some research effort has been concentrated on clarification of the mechanism of vocal-cord vibration. Van den Berg (1958) first proposed a myoelastic-aerodynamic theory, which established a physical interpretation that is now considered standard. J . L. Flanagan, K. N. Stevens, and K. Ishizaka have contributed improvements and data, as well as deeper insights into the mechanism and its linguistic implications (Flanagan 1965; Flanagan and Cherry 1969; Ishizaka and Matsudaira 1968; Halle and Stevens 1967; Stevens 1971).

The main point of interest here again is to see which part of the voice modulation in speech can be attributed to laryngeal control gestures and which to a more physical sort of effect caused by other kinds of control, such as articulatory closure. The role of the pulmonary contraction gestures in relation to segmental and non-segmental properties is also not understood though there have been warm disputes about it.

By recent fiberscopic observations of the larynx during speech utterances, it has been clarified that tense or voiceless stops in Japanese and English exhibit laryngeal gestures that very widely depending on the

phonological environment (Sawashima 1968, 1970; Sawashima et al. 1970; Fujimura and Sawashima 1971). Even though the glottis is kept wide open during the closure period for these consonants in absolute initial prestress position, the same consonant in word medial position is produced with an almost closed glottis. In the latter case, the slit is considerably wider for voiceless fricatives in the same environment. This may be interpreted as a purely physical or perhaps partially physiological (feedback dependent) consequence of the same laryngeal control and different flow conditions for the two different classes of consonants, or perhaps as inherently different motor specifications, depending on whether the consonant is a stop or a fricative. When we do not know the phonetic fact precisely, linguistic studies of languages do not tell us definitely, on independent grounds, what the phonological units and their possible values are, or in particular whether there is one dimension or more for the glottal constrictive gestures. Stevens and Halle discussed the possible relevance of a particular physical dimension--the coupling constant for two masses in Ishizaka's vocal cord model--to this important question in phonological theory (Halle and Stevens 1971). Ishizaka's two-mass model of the myoelastic-aerodynamic vibration system was proposed in computer simulation studies as a technical improvement over Flanagan's model that used one vibrating mass for the vocal cord (Flanagan 1965, Ishizaka and Matsudaira 1968). But from Stevens-Halle's point of view, the additional degree of parametric control of the vocal-cord vibration could be crucial for phonetic interpretation of the linguistic sound system. Whether this particular proposal is true or not is not known, but it will not be difficult to see that this is a physical or physiological matter that interacts with fundamental phonological interests. In a sense, a physical or physiological model can be gross or must be detailed, depending on whether the detail is used in language for a functional distinction. What kind of details can be utilized--or rather what are details and what are gross facts--is a fundamental and non-trivial biological question that relates to the essential nature of human cognitive behavior. K. N. Stevens points out a possible principle as a criterion of this issue: physical correlates at the consequent signal level must be little affected by deviations from the standard in motor gestures for the latter to be used in language

as physiological correlates of a phonological unit. He suggests that this stability principle works particularly well when the mapping between the control and output levels is non-linear (Stevens 1971 and to be published).

1. 3. 2. Pitch and Intensity

In the case of voiced sounds the acoustic waveform is quasi-periodic, and the fundamental period is determined by the rate of vibration of the vocal cords. The fundamental frequency observed in the harmonic structure of the spectrum generally changes during an utterance, and the changing pattern reflects physiological control as well as natural tendency as a consequence of, for example, the decaying pulmonic overpressure towards the end of a stretch of voicing. Both the so-called subglottal pressure, which reflects the pulmonary state and pertinent respiratory efforts, and the laryngeal gestures affect the voice fundamental frequency, or pitch as it is often called (with a psychoacoustic connotation). Which physiological control dimension corresponds to which linguistic or non-linguistic feature is an important and largely unanswered question, except for some few well-established functional aspects of the laryngeal muscles, in particular the activity of the cricothyroid in relation to the accent patterns of Japanese and Swedish (Öhman et al. 1967; Gårding et al. 1970; Simada and Hirose 1970, 1971).

Some characteristics of the pitch contour are used for identification of various phonological units, i. e. they manifest sentence types (intonations), phonological phrase boundaries (the configurational features in Jakobson et. al. 1951), prosodic (suprasegmental distinctive) features, and also inherent (segmental) features. For the last aspect, which is perhaps less recognized, there is experimental evidence that the pitch inflection pattern near the stop release in a stop-vowel syllable serves as a secondary cue for the perception of the tense-lax or voiceless-voiced distinction of the consonant, while the onset of voicing in reference to the explosion plays the role of the primary cue (Fujimura 1971). In analyses of natural utterances, too, this effect of consonants on the pitch contour cannot be ignored, when we wish to account for the actual time course quantitatively.

Correlation between the voice intensity and fundamental frequency is an

intriguing question.* It used to be held, traditionally, that English is a "stress language" whereas Japanese is a "pitch language." This is incorrect, at least in the sense that one can produce rather natural sounding synthetic speech in English without distinctive use of intensity at all. The concept of "stress," referring to (necessarily subjective) intensity, in opposition to "pitch," is a dubious notion itself, even though it may sound physically simple and clearcut. The physically defined intensity is so severely affected by the articulatory condition, that it can scarcely be correlated with so-called stress. The source intensity for voiced segments, ignoring some effects of the vocal tract-source interaction, makes sense, but this quantity is not easily measured from speech waves. At the moment, we seem to lack reliable data about the correlation between intensity and pitch.

Extraction of the fundamental frequency is not technically simple either, but this problem has been studied intensively for a long time.** This is another of the cases in speech research where a problem looks simple but is technically extremely difficult. At present we have a few sophisticated enough methods to solve the difficulty. For phonetic studies the most effective method probably is the (narrow-band) spectrographic examination.

2. Observations of the Natural Process of Production

As the basis of hypothesizing a working model of the human speech production process and also for quantitative data that we may match against the consequences of the hypothesized model for its evaluation, we need an extensive body of accurate data on natural articulatory movements and laryngeal gestures. To gather these data we have developed several new techniques which may be worth mentioning here.

---

* See Lehiste 1970 and Ladefoged 1967 for a comprehensive review and relevant discussions.

** See Flanagan 1965 for a review.

2. 1. Articulatory Movements

In the classical phonetics literature the phonetic values were described almost exclusively in terms of articulatory conditions. After intensive studies of the acoustical structure of speech signal for ten to twenty years, our recent attention is drawn back to the old topics, but to gather accurate and quantitative measurements from a more extensive corpus of utterances, rather than to make occasional qualitative and subjective observations as was so often done in the past.

2. 1. 1. Optical Observations

Visual observation of articulatory movements has been, perhaps next to auditory evaluation, the most common technique of phonetics, even though quite often it was not mentioned explicitly in the literature. It is still one of the most useful research techniques, particularly when supplemented by modern devices such as high speed motion pictures, fiberoptics, high-sensitivity photo-multipliers, etc., and perhaps even more so when used in combination with an interactive computer for data processing.

It seemed that the effective use of high speed photography for studying articulation had been somehow ignored until about twelve years ago, when I designed an experiment at MIT specifically for collecting quantitative data of the speech dynamics of the lips and the mandible (Fujimura 1961a, 1961b). Attention was directed in particular to separating the physiologically controlled gestures from the uncontrolled movements. For example, it has been found that there is an uncontrolled component in the motion of the lips immediately after the explosion of an initial stop consonant--a so-called ballistic motion. In connection with this, the labial movement in this kind of environment was found to be so fast that there would be no meaning in expecting direct perceptual effects from formant transitions as such in the initial part of the consonant-vowel transition.

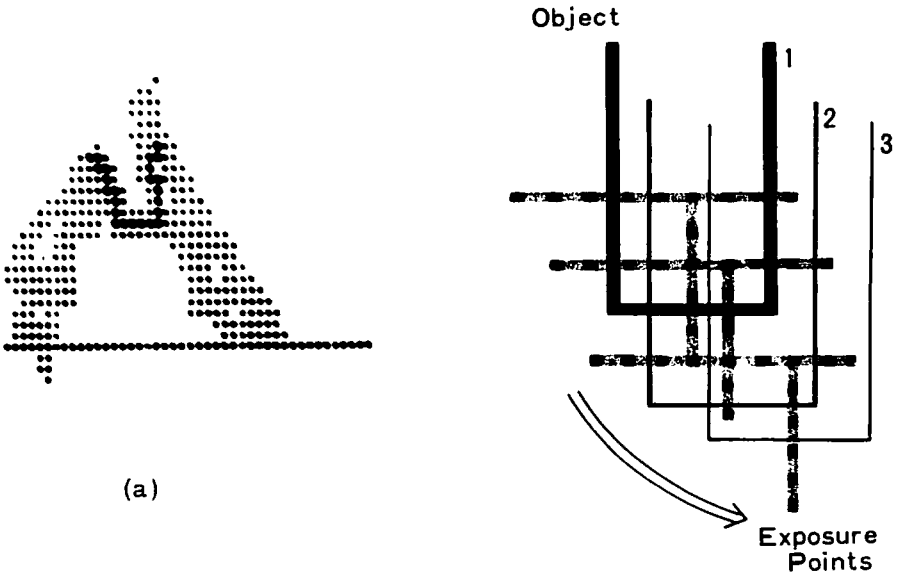
A moderately high frame-rate--like a few hundred a second--combined with the use of a modern fast stroboscopic technique is sufficient for the purposes of these studies. Since optical measurement gives accurate data without interfering with the articulatory actions, it is substantially superior to other techniques that require mechanical measurement. The main

weakness of this approach, apart from the obvious point that we cannot observe hidden articulatory organs like the tongue, is that data processing is tedious and time-consuming. Some techniques have been devised in order to record the data directly as curves representing important articulatory variables, but, in general, direct observation of the body surface is not only more informative but also quite often necessary in order to avoid gross errors of measurement. Moreover, direct observation is very useful for gaining insights in the first stage of the exploration to determine the most relevant factors to be detected and measured in the processing of the optical data. Computer processing of optical images with advanced techniques of pattern recognition, or computer-controlled object-search techniques, may be the best solution (cf. the radiographic technique, infra) to the data processing problem.

2. 1. 2. Radiography

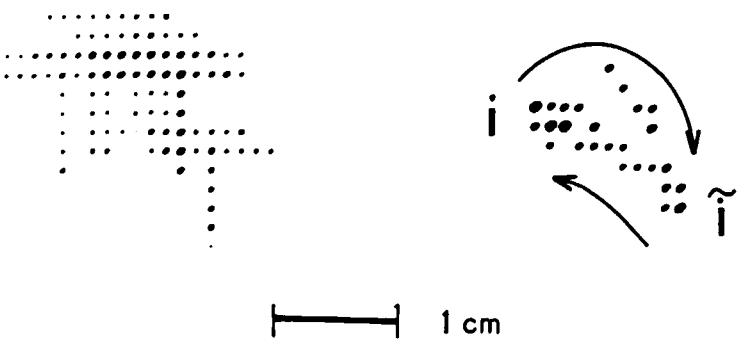
For a long time, radiography of various sorts, including in particular high-speed cineradiography, has been the major source of objective information about the articulatory behavior of the tongue. Unfortunately, the hazardous effects of radiation on the subject limit us when we wish to use these methods, and, as a result, the amount of data is always unsatisfactory from the point of view we discussed before (cf. Introduction). Deriving data from the cineradiographic images, in addition, is not only tedious, as in the case of optical images, but also simply not feasible by any automatic means.

A new technique is being developed at our laboratory in order to solve both problems at the same time (Fujimura et al. 1968 and to be published). It employs a computer to control the direction of a fine x-ray beam, which after passing through the object is detected by a high-sensitivity scintillation counter. The radiopacity at the selected point (about 1 mm in diameter) in the object is thus estimated and read into the central processor in digital form. This method, when technically well developed, will provide us with means to obtain useful information about tongue movements as physiological correlates of different kinds of phonological units. At present,



(a)

(b)



(c) (d)

Fig. 1 (a) An outline of the identified fiberscope (thicker dots). The thinner dots represent the entire set of examined sample points during this object recognition process. (b) Sample points to be exposed for determination of the position of the fiberscope during the tracking of its movement. Three successive "frames" are shown. (c) T-shaped arrays of the exposure points over 15 frames during an [i] to [$\tilde{i}$] transition. (d) The track of the position of the tip of the fiberscope automatically determined in real time over a 1-sec time interval during the test utterance of [i: $\tilde{i}$: i:].

however, a device exists only for pilot studies of the new method itself. Even though preliminary results have been very successful (Kiritani and Fujimura 1970, Kiritani 1971), the acquisition of actual data in speech research remains in the future until the completion of a larger x-ray device. One modest exception is its use in actual real-time monitoring of the tip of the fiberscope which is inserted into the pharynx for observation of the glottis during speech utterances. Fig. 1 illustrates an example of the automatically located fiberscope, the points of x-ray exposures for the automatic identification and tracking, and a track of movement during velar articulations.

Another limitation on measurements by lateral cineradiography is the difficulty of estimating the three-dimensional shape of the organs that surround the vocal tract. The consequence is that we must guess about some of the cross-sectional areas along the vocal tract axis needed for calculating the acoustic consequences such as formant frequencies (see §3). Also, for physiological considerations which can be crucial for determining what the effective physical dimensions are for the dynamic characteristics of the articulatory movements, particularly of the tongue, quite often we wish to determine the movements of specific points fixed on the tongue rather than the surface contour of the tongue. This information can be obtained by placing small metal pellets on the tongue surface, which at the same time makes it feasible to measure the tongue position accurately in each frame. It can be advantageous, to some extent, also in estimating the three-dimensional shape of the tongue. For these reasons, this pellet technique has been employed in some recent representative cineradiographic studies (Houde 1968a, 1968b; Perkell 1969; Kent 1970), and will be used in the computer-controlled microbeam method, too. Tomographic methods of various kinds can be employed for some specific problems (Hollien 1965), but they suffer from the dosage limitation even more seriously than regular cineradiography.

2. 1. 3. Palatography

Another old technique in experimental phonetics is palatography, which we must mention here. Its obvious shortcoming is that the palato-

gram records the total area--that is, the greatest extent--of palato-lingual contact over the entire time course of the test sample, and no information can be obtained on the movements of the tongue. Recent techniques developed independently by a few groups in different parts of the world give more or less similar results by recording the temporal changes in lingua-palatal contact by use of electric conduction through the tongue surface (Rome 1964, Kozhevnikov and Chistovich 1965, Shibata 1968, Hardcastle 1969). The dynamic palatography developed in our laboratory typically employs 64 sample points on a thin artificial palate fabricated for the particular subject, and the data can be recorded in two different forms. One output form is the "palatospectrogram" which displays, together with part of the regular sound spectrogram, time varying on-off traces of the lingua-palatal contact for each of the 64 electrode positions, which are assigned 64 different frequencies (Shibata 1968). This is a convenient way of recording the data for visual inspection. The other output is a computer-mediated oscilloscope display in the form of a slow-motion movie of the palatal pattern frames (Fujii et al. 1971, Fujimura et al. 1972). Figure 2 illustrates an example of this display, which includes a curve representing the speech amplitude; the time of the pertinent frame can be identified as a brightened dot on the speech amplitude curve. Both the palatal (automatically dichotomized signals and the (analog) speech signal can be stored in real time on magnetic tape by a special analog-digital hybrid recorder (Ishida 1969). This method is particularly suitable for automatic digital processing of the palatal data by a computer, when due care is taken. Palatography does not seem to cause any biological disturbance and it is practical therefore to collect and analyze a very large body of data. Variability of speech utterances can be quantitatively estimated by this technique, and some of the results are being published (Fujimura et al. 1972).

Even though the palatal contact patterns as time series do not provide us with complete information about the tongue movements, data in this form will be particularly useful in complementing radiographic data which are incomplete by themselves. For example, the variability to be observed among repeated utterances of the same word or sentence has rarely been studied quantitatively in most aspects of speech production (but see Malecot

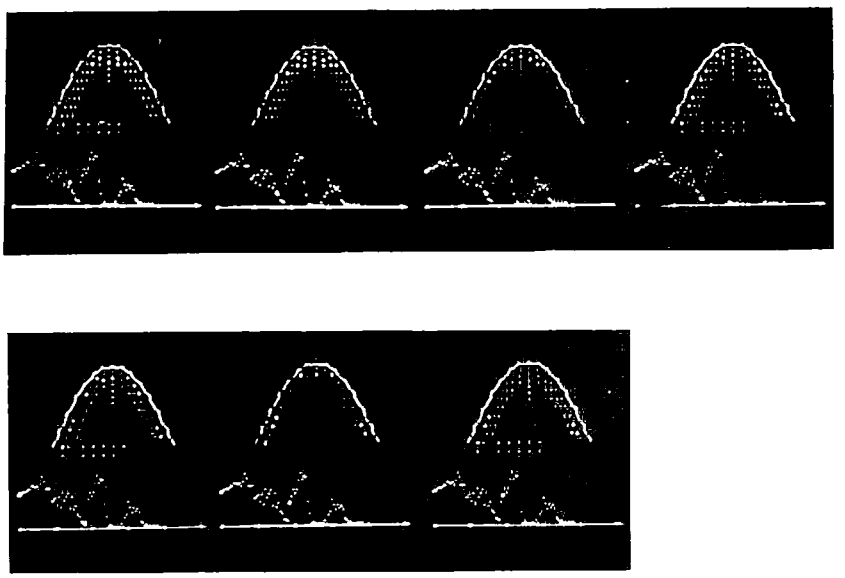


Fig. 2 A series of consecutive display frames, showing the change of articulation of [r] in the Japanese sentence [mo:araka]. The brightened spots inside the horseshoe shape, which represents the shape of the palate, indicate the positions on the palate where the tongue is in contact with the palate. The lower curve represents the speech signal envelope, and the brightened dot indicates the time pertinent to the frame. Frames are 10 msec apart in time. The frame preceding this series did not show any contact of the tongue.

1968, Lubker and Parris 1970). Variability in respect to the phonological context or speech environment is, needless to say, a point of common interest for both basic research and applications, but for obvious reasons past treatments even on the acoustic level have tended to be restricted in the size of available, or rather processable, material. Such basic data, particularly those describing articulatory movements, will be important in making a substantial step forward in research dealing with automatic speech recognition or speaker identification.

2.2 Laryngeal Gestures

One of the least understood topics in experimental phonetics is how one controls the laryngeal conditions in speech, particularly in connection with the production of various consonants. For example, a gross qualitative statement as to whether the glottis is open or closed during the articulatory closure of the non-aspirated stops in French and many other languages has been simply a matter of conjecture. Observation of the glottis during real utterances with an actual stop of air flow--an impossibility for the standard laryngoscopic techniques--is necessary to understand the essential characteristics of the tense-lax and/or voiced voiceless distinctions in different languages. The interaction between the laryngeal and the vocal-tract conditions through the aerodynamic process is so strong that we cannot simulate the real situation during an articulatory closure with the air passage of the vocal tract left open. Radiographic studies have been attempted in this area, too, but the radiation dosage problem is especially severe here. In particular, when one attempts to take frontal views of the larynx for studying the glottal adduction-abduction process, we cannot avoid including the radiosensitive vertebrae in the image field. Also, the structures forming the glottis are not favorably radiopaque, and any effort to reenforce the radiopacity artificially tends to disturb the mechanically and sensorily delicate glottal actions in natural utterances. Within these limitations, however, there has been some interesting work of linguistic relevance. For example, Kim in his studies of Korean stops, advanced an account of the degrees of aspiration in terms of the glottal conditions at the moment of articulatory release (Kim 1970). Some other studies are particularly concerned with the role of the pharyngeal conditions in relation with the source or tenseness feature. Perkell, for example, concluded in his study of lateral cineradiographic images that English lax stops were characterized by more yielding pharyngeal walls than tense stops. Whether this effect is caused by the air pressure against a passively compliant tongue surface with the compliance being dependent on the muscular states, or is the result of active motoric enlargement of the cavity for facilitating the vocal-cord vibration is a matter for future study (Perkell 1969).

Recently, the gaps in our knowledge of glottal conditions are being filled by data obtained with a new optical method (Sawashima and Hirose 1968, Sawashima and Ushijima 1971) utilizing a fiberscope which has been specially designed for laryngeal observations. A flexible cable 4 to 6 mm in diameter contains two bundles of glass fibers; one is a coherently arranged fiberoptic bundle for image conduction, and the other for conducting an illuminating light. The cable, the tip of which houses an objective lens, is inserted through the nasal passage down to the middle pharynx, and the image of the glottis is viewed or photographed through an eye piece attached to the outside end of the cable. With an appropriate light source connected to another branch of the cable, a color movie readily can be made, typically with a rate of 60 frames /sec. In this way, the overall laryngeal state in segmental articulations can be directly viewed and recorded. The pitch control gestures also can be studied by examining the qualitative changes in the viewed vocal cords during vibration, and some consistent up and down movements of the larynx associated with pitch changes in speech utterances also have been observed.

For the laryngeal gestures in connection with the problem of manner distinctions or their relevant source features, a particularly interesting case is the Korean stops. It has been suspected, from some cross-language studies on the so-called voice-onset time (referred to the articulatory explosion), that the phonetically significant physical dimensions, or the corresponding physiological correlates, in Korean might be substantially different from many other languages (Lisker and Abramson 1964). Some preliminary results of fiberscopic studies have provided very interesting facts abo t this question (Kagaya 1971). Some examples of the laryngeal images for the three manners of the Korean dental stops are illustrated in Fig. 3.

Among the three types of dental stops,* designated by /T/, /t/, and

---

* For some relevant studies of the Korean stops, see, among other studies, the following works of acoustical analyses: Umeda and Umeda 1965, Kim 1965, and Han and Wietzman 1970.

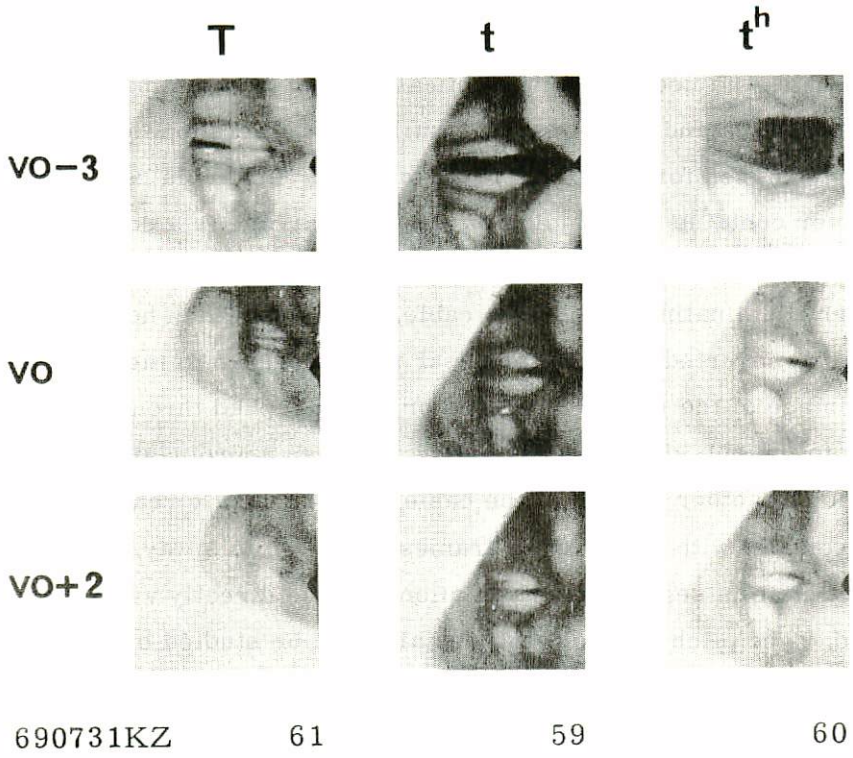


Fig. 3 The glottal images viewed through a fiberscope during the articulations of the three manners of dental stops in Korean. The upper row gives the frame three frames (about 120 msec) preceding the voice onset for each of the utterances, the middle row at the voice onset following the absolute initial stop, and the bottom row two frames (about 80 msec) after the voice onset for the following vowel /e/.

/t^h/ in Fig. 3, respectively, the first is called the forced type, or the tense non-aspirated, and it is known to be always voiceless during the articulatory closure. The "tenseness" of this type of articulatory manner (Kim 1965) may be related to the following observational peculiarities in the case of isolated /CV/ syllables (Kagaya ibid.): (1) the vocal cords, particularly the vocal processes, become closed more than 100 msec prior to the voice onset, which takes place at the time of articulatory explosion; (2) the closing action of the vocal cords is quite rapid; (3) the larynx is sharply lowered at the same time as the glottal adduction. The

pitch when voice is initiated, however, is markedly high compared to the other two types in the same environment, which show apparently higher laryngeal positions. This is unusual when compared to the usual correspondence between a higher pitch and a higher laryngeal position.

The "lax" stop, /t/ in Fig. 3, shows a slow closing action of the vocal cords, and voice onset is slightly later than the release of the articulatory closure, leaving some short length of aspiration. Even after the voice onset, in this case as well as in the heavily aspirated type /t^h/, the glottis at the vocal processes is still slightly open. The lax /t/, incidentally, becomes voiced all through the articulatory closure when the consonant is in word-medial intervocalic position.

Even though it is too early to draw conslusions from the limited data of one subject, * it may be worth while to hypothesize as follows. In the case of the lax articulation, the vocal cords are loosely set to a position for vibration with a slack gesture (Halle and Stevens 1967) somewhat ahead of the articulatory explosion. As soon as the air flow builds up beyond a threshold value after the release of the articulatory stop, the cords start vibrating. The forced type, in contrast, has a positive adductive gesture well in advance of the explosion. The pressure difference across the glottis, which is necessary for the vocal cord vibration, may be created by a positive lowering action of the larynx, expanding the supraglottal cavity (Fischer-Jørgensen 1963, 1968). As the vocal cords are tightly closed, they block the air flow up to a certain value of transglottal pressure difference, and this blockage will be broken at the moment of articulatory explosion. There may be an appreciable amount of acoustic interaction between the articulatory stop release and initiation of vibration of the stiff vocal cords. The stiffened vocal cords and the transient transglottal pressure cause the observed high pitch. The third type, the heavily aspirated, is characterized by a wide opening of the glottis which is maintained well through the moment of articulatory stop release, in conformity

---

* We have some supporting data of another native speaker in the form of optical glottographic records.

with Kim's account (Kim 1970). The vocal cords may be stiff or slack, and when they are drawn together for the succeeding vowel gesture they are ready to vibrate even before the adduction of the arytenoidal cartilages has been completed, because the air flow rate is already high.

If these hypothetical characterizations of the source features for the Korean stops are correct, we should be able to observe a marked transient drop of the supraglottal pressure immediately before the explosion of the force-type stop, and perhaps also its reflection on the apparent state of the articulator, e. g. the external shape of the lips for [p], as was observed in studies of American English stops in contrast to [m] (Fujimura 1961b). The forced stop is also expected to show a marked activity of the vocalis muscle. * We should be able to simulate the characteristic transient phenomena of these different manners of production by an appropriate computational model, if we couple the already available vocal-cord vibration model with an appropriate vocal-tract simulation (Flanagan and Cherry 1969, Flanagan et al. 1970).

We may go a step further and hypothesize abstract phonological specifications for the Korean stops, in convormity with the above-mentioned tentative account of the phonetic facts. Let us assume that there are distinctive features in respect to the laryngeal gestures, adduction-abduction and stiff-slack, roughly along the lines proposed by Halle and Stevens (1971). For the time being, let us adopt the common assumption that the two features are assigned one of binary values, and when a feature value is not specified at the abstract level, it is later determined by some context-sensitive (assimilation) rule of Korean phonology. ** Let us assume, in particular, that the lax stop is assigned no specification for adduction or abduction, but is assigned a specification for the slack-stiff dimension.

---

* Marked electromyographic activity of the vocalis muscle in connection with glottal stops, as well as some other related problems of source control, are discussed elsewhere (Gårding et al. 1970).

** As a relevant example, see Chomsky and Halle (1968) for a detailed proposal of the descriptive framework of a phonological rule system.

When there is a certain kind of boundary preceding the pertinent consonant segment, then this segment will be given the status abducted. In other words, the boundary as a phonological unit may be considered to have an assignment of abduction status, and an assimilation (redundancy) rule will assign an abducted status to the contiguous consonantal segment that is non-specified with respect to this feature. There is some experimental evidence in the data referred to above, however, that the closing action toward the lax voiceless stop is loose and unstable. This may be taken as an indication that the assignment of the abducted status for the stop segment is not made discretely, as either plus or minus by a phonological rule, but the non-specified status is carried down to the phonetic level where coarticulatory rules process continuous time functions of a physical nature.

The Korean case, of course, has important bearings on general phonetic theory as a crucial special case, and the hypothesis proposed above does not agree with the account proposed by Chomsky and Halle (1968) nor with a later tentative theory by Halle and Stevens (1971).

In another interesting series of studies using the fiberscope, Sawashima and his coworkers have clarified the characteristics of the vowel devoicing phenomena in Japanese (Sawashima 1971, Sawashima et al. 1971). Comparing, among other things, a set of words like /sekikei/ and /seQkei/, which are actualized as [sekɨke:] and [sekke:], respectively, the width of the glottal opening, both at the vocal processes of the arytenoids and the membranous portion of the cords, was measured as a function of time. It was found that the word-medial sequences [kɨk] and [kɨt] had significantly larger glottal openings than the geminate consonants [kk] and [tt](Fig. 4) even though the durations of these two kinds of segment sequences were approximately the same. It was concluded, based on this finding, that the glottal maneuver for devoicing the vowel is not a mere skipping of the phonatory adjustment for the vowel, but a positive gesture of glottal abduction, even though this devoicing is not phonemically distinct. * This is probably one of the many allophonic rules that assign specific feature

* Supporting EMG data have been discussed by Hirose (1971).

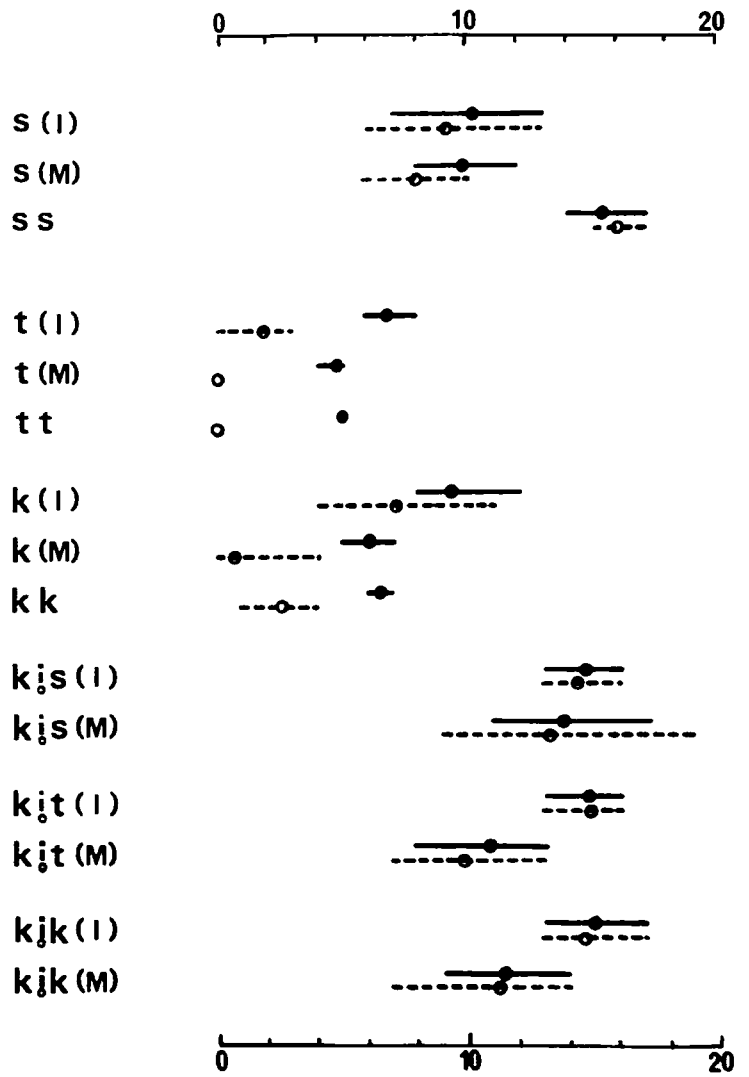


Fig. 4 Peak values of glottal width for voiceless (sequences of) segments. Filled circles and solid line segments represent measurements at the membranous portion of the vocal cords (the maximum distance), and the open circles and dotted line segments those at the vocal processes. The circles are for the mean values, the line segments the entire ranges of observed values. There were 59 samples for /s/ in word initial position (I), 18 and 12 for /k/ and /t/ respectively in word medial position (M), and 6 for each of all the rest. The values are in an arbitrary scale.

values depending on the context prior to the application of a coarticulatory rule of a physical sort.

The fact that voiced consonants are usually associated with a lower pitch (see § 1. 3. 2), might be a result of a particular laryngeal gesture for voiced consonants (Chomsky and Halle 1968); at present this suspicion is but a matter of conjecture. Pitch control is usually associated with some change in the height of the larynx, at least for the accent pattern in Japanese** and the tone in Chinese (Chuang et. al. 1971). In the case of Japanese, the pitch control can be observed in respect to this physiological correlate, even in devoiced vowel segments (Fujimura 1971J).

3. Functional Models

An ideal methodology for studying the production mechanism might be as follows. Data should be derived from natural utterances at different levels of production, such as cerebral motor commands, neuromuscular activities, mechanomuscular states, proprioceptive and other sensory afferent neural feedback, states (shapes) and movements of the speech organs, vocal tract dimensions, the area function (see infra), and formant and antiformant frequencies. Similar specifications would have to be obtained for source characterizations, such as laryngeal gestures and pulmonic conditions and aerodynamic states near and above the glottis. All these combined would enable us to describe the course of events from linguistic code specifications to the sound wave, by proposing theories and models to account for the relation between contiguous levels, and checking level by level with observed data from natural utterances. Obviously, the state of our studies is far from this ideal case.

As we have seen above, of course, experimental techniques for observing and measuring the natural processes of speech are often very new, and we anticipate interesting data to come out of the experiments in

---

** For some illustrative examples, see "The Larynx in Speech Utterance," Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo (demonstration, 16 mm in color).

the near future. No doubt these results will enhance our knowledge of the speech production mechanism at more than one level of the hierarchical process. Yet in view of the inherent complexity and variability of the phenomena under discussion, it is to be doubted whether these empirical observations alone will lead us to usefully organized and comprehensive descriptions of the natural processes. Even though the physical constraints of the production mechanism are the major basis for use in disambiguating the complex phenomena, the anatomical structure of the speech organs is formidably complex, and the physiology of no single organ, not only the tongue, but also the lips or even the rigid bone of the mandible, is known thoroughly for the intricate speech gestures. Rather than going into deatils about these problems, I would like to briefly go over the problems we have to face in relating the mechano-physiological findings to the acoustic phenomena.

Since we do not have sufficient knowledge about the higher (or rather intermediate) levels that constitute part of the speech production mechanism, we will try to compensate for these gaps by constructing a model for the lower levels and piece together fragmental data at different levels in order to check the adequacy of the hypothesized model. From this point of view, our theoretical and observational data by and large are firm for the acoustic phenomena, and the main point of the present research efforts is to provide data for constructing a workable dynamic model of the mechanophysiological articulatory system, a model which will take input specifications that are more-or-less directly interpretable in terms of discrete linguistic codes.

3.1 Specification of Area Functions

One level of description we have to work out within this framework is a quantitative specification of vocal-tract area functions. These functions will be determined by the states and movements of the speech organs, and will determine the acoustic characteristics of the speech phenomena. Some early work offered simple and effective mathematical functions that are characterized through a set of three "articulatory" variables, which represented, approximately, the place of the main lingual constriction, the extent of that lingual constriction, and an acoustically effective measure of the labial constriction (Stevens and House 1955, 1956; Fant 1960). Based

on this descriptive framework of the area function, some useful charts were provided by these workers relating these articulatory variables to the lower formant frequencies. Considerable non-linearity in the relation was apparent, which carried important implications for the succeeding work.

In order to substantiate the validity of this kind of parametric specification of the area functions, we need more area functions for natural utterances. Unfortunately there has been little success in estimating the area function by direct measurements. * Radiographic studies, which are pertinent to the shape of speech organs, a higher level description, are at present the major source of information for deriving the area function. A difficulty arises, from this point of view, in relating what we can obtain by radiographic measurements to the area function.

Let us assume, to be concrete, that we have good and accurate estimates of the midsagittal contour of the tongue surface obtained from the lateral x-ray frames. At the same time, we would have good simultaneous sound recording for the utterances. As we have discussed in § 1, we have for most purposes firm enough theoretical means to numerically derive formant frequencies for the given area function. On the other hand, we have reliable means to derive formant frequencies from the speech waveform, i. e. the formant extraction techniques (Flanagan 1965 , Schafer and Rabiner 1970, Olive 1971). Thus we are able to test the adequacy of our theory and the reliability of our measurements by comparing the data that are obtained separately for the same utterance at two distinct levels, i. e. the x-ray measurement and the acoustic waves, and evaluate the match between the derived sets of formant frequencies.

---

* There is still a hope to measure the acoustic impedance looking into the vocal tract at the mouth opening, which in addition to the series of formant frequencies, gives another series of characteristic values, viz. the zeros of admittance. These would serve as the mathematically required information for determining the acoustically effective area function (Schroeder, 1967, Sondhi and Gopinath 1971).

The main problem then is the conversion from the midsagittal cross-dimensions to the acoustically effective cross-sectional areas for the formant calculation. Research workers have been concerned with this problem for the past several years (Fant 1960 , Heinz and Stevens 1965, Sundberg 1969). Obviously, we need a simplification in the form of a three-dimensional model of the deformable speech apparata. Simplified parametric descriptions and related computations tend to be still quite complex, but appropriate techniques of analysis-by-synthesis seem to work effectively. Consistent and plausible descriptions of the articulatory conditions have been derived for some simultaneously recorded cineradiographic and acoustic data of vowels (Maeda 1971).

It is worthwhile at this point to raise the question whether a more straightforward approach can be applied to the problem at hand. If we could make an inverse calculation that derives the area function directly from the formant frequencies, or from any other information available in the acoustic speech signal, then we would have for comparison the x-ray measurements of the midsagittal vocal tract dimensions on the one hand and the area function thus derived on the other. By accumulating data for different articulatory conditions (let us say for the same subject), we would be able to find quantitative relations between the midsagittal linear dimensions and the areas for different cross sections along the vocal tract, without any indirect analysis by synthesis. Many efforts have been made to clarify the problems pertinent to this issue, and very interesting results have been provided through theory and computations (Mermelstein 1967, Schroeder 1967). In my opinion, however, the particular point at issue has been concluded negatively. If we consider a lossless acoustic system for the inverse calculation from formant frequencies to area function as a general problem, we can show theoretically that the solution is infinitely ambiguous. In other words, the set of formant frequencies simply does not provide us with sufficient information to determine the area function. Measurement of the formant bandwidths in speech spectra does not lead to any satisfactory solution either.

The acoustic theory and synthesis experiments based thereon tell us that a specification of the three formant frequencies suffices for determina-

tion of the perceptual phonetic quality of any vowel (Delattre et al. 1952, Fujimura 1967). It could be said, therefore, that there are only three degrees of freedom for the independent variables to describe the overall articulatory conditions of vowels, if we are concerned strictly with the spectral quality of vowels. Consequently we should be able to determine these articulatory parameters--instead of the area function with too many unknown variables--from measured formant frequencies. Our point here is however, not in such "articulatory" representation of the acoustical quality, which could possibly be useful in applications like speech bandwidth compression systems.

If we consider the human physiological capability in the universal phonetic sense, we may well assume that there are more than three muscles which can be controlled independently for the generation of different vowels. Even for the labial gestures alone, it is actually observed that protrusion with rounding is not necessary for some significant labial constriction, as seen in the distinction between /ʉ/ and /y/ in Swedish. Whether these two vowels represent a phonetic minimal distinction in this respect may be somewhat unclear (Fant 1971), but in any case it seems hard to argue with our present pertinent knowledge of physiology that this labial distinction must necessarily be coupled with some difference in the lingual gesture. The situation is not essentially different for the lingual gestures proper. It is reasonable to assume that for any of the commonly known languages the number of independent muscular controls for the lingual states that are utilized in vowel articulation is more than three (e. g. the anterior and posterior portions of the genioglossus, the styloglossus, and at least one more for the mandible position).

While the anatomy of the human speech apparatus restricts the set of possible area functions for the vocal tract, it is at the same time likely that not all the physiologically independent muscular controls are effectively independent in determining the acoustically relevant specification of a given area function. The labial articulations in Swedish vowels exemplify the case. Whatever the articulatory features may be, the acoustical effect of labial constrictions can be approximated by a lumped mass of the air at the orifice and can be represented by a single measure, viz. the effective

opening area divided by the effective length of the orifice (Stevens and House 1955). The issue is then not really the number of independent variables in articulatory control of vowel quality, but in the lack of correspondence between the articulatory and acoustical levels of description. What then, are the most effective articulatory dimensions as physiologically controlled and linguistically relevant variables? In inquiring into this, it turns out that we have to consider not only static articulations of vowels, but also the dynamic aspects of speech production, the temporal organization.

3.2. The Cylinder Model of the Tongue

The major portion of the midsagittal outline of the tongue in various vowel articulations can be roughly represented by a circle with a fixed radius but movable in two dimensions--front-back and open-close--relative to another circular fixed wall which represents the palate and the posterior pharyngeal wall. The gap between the two circles (with a straight portion appended near the glottal end), together with an adjoining lip section, forms the vocal tract. I proposed this rough static model and tested it informally by use of an electrical vocal tract analog at MIT in 1959, and my intention then was to see how we should relate the dichotomous distinctive feature specifications of vowels to articulatory specifications and then automatically to acoustic signals. Selecting appropriate values for the radii of the circles and also adopting non-orthogonal geometrical axes for the two articulatory dimensions, open-close and front-back, the model seemed to work grossly, even with the crude assumption of a proportionality between the gap dimension and the cross-sectional area. Later at Bell Labs, Cecil Coker adopted this model as a starting point and elaborated it into a new dynamic model with many novel and interesting features, some of which seem to point to very essential characteristics of the temporal organization of speech (Coker and Fujimura 1966 , Coker 1968, Glanagan et al. 1970). The model has been implemented as a computer simulation program in combination with a hardware terminal-analog synthesizer and is available for a rather large-scaled experiment of synthesis-by-rule. Coker jointly with Mrs. Noriko Umeda, a linguist by training,

has developed an elaborate, and in a sense amazingly complete, system which approximates the entire chain of processes from linguistic codes to the sound waves.

Some workers recently have advanced more physiologically motivated models, particularly in explicit reference to the role of the mandible in determining both the lingual and the labial constrictions (Lindblom 1965, Mermelstein et al. 1971). In a comprehensive report Lindblom and Sundberg (1971) also discussed anew based on this descriptive model the old problem of acoustical interpretation of the articulatory variables. The classical notion of tongue height is formally analysed into two essentially different components, one due to the mandible position and the other reflecting efforts of the tongue muscalature. The same effect is claimed by Mermelstein and his coworkers, who proposed mathematically parameterized tongue shape specifications in reference to the mandible base. Both of these models, as well as the cylinder model, try to represent the apparent restriction imposed on the laterally observed tongue surface contours for different vowels, as discussed in some detail by Kent in his cineradiographic studies (Kent 1970). These models take note of the fact that the vocal tract is bent, which is essential in consideration of the occurrence of articulatory constrictions caused by different locations of the tongue body, but not relevant to acoustical considerations.

3.3 Temporal Organization

The Coker model is effective as a research tool particularly because the transitional characteristics can be described more simply on the articulatory level than on the acoustical level--presumably due to the fact that speech dynamics are determined essentially by articulatory physiology rather than auditory physiology. It has long been known that various physiological variables, or their pertinent mechanical sybsystems, not only have time constants of their own, but also are often out of synchrony (Fujimura 1961a). This situation makes it difficult to discuss the temporal structure of speech fully by examining the acoustic phenomena alone, even though it is to be acknowledged that many valuable facts have been deduced at this level of description (e. g. Delattre et al. 1955, Lindblom 1963,

Öhman 1966, Stevens et al. 1966). The effects of controls in all of these dimensions are collapsed into one and the same acoustic variable, the sound pressure, even through the highly nonlinear mapping as mentioned before.

We may assume a feature matrix model for the sequence of codes that specify speech gestures for a given utterance unit (Jakobson et al. 1951; Fant 1962, 1971; Chomsky and Halle 1968). This model is basically segmental in the sense that all temporal information is represented by a horizontal arrangement of columns, each of which correspond to, roughly speaking, the phoneme. Details of temporal organization for phonetic degrees of freedom are specified through the mechanism comprised of phonological rules and phonetic actualization processes. The input of this interpretive transducer may be taken as the syntactic surface structure of a sentence (or some other linguistic form); the output will be the phonetic event, as represented at some level of physical or physiological description of the speech production process (Fujimura 1967J). The information carried by the output, however, does not correspond exactly to that at the input level. Extra-linguistic information such as expressive features (Jakobson et al. 1951), voice characteristics, etc., must be added to account for actual human speech phenomena. Also, some additional modifications will have to be treated as random factors.

Within the realm of linguistic problems, we still need some procedure for phonetic realization in order to complete our theory of sound shape up to the level of empirical tests. Just like the notion of kernel sentences in the earlier formulation of transformational syntactic theory (Chomsky 1957), we may propose a minimal phonetic realization procedure in order to derive a "standard" phonetic event for a given linguistic specification of an utterance (Fujimura 1970). The production of "kernel utterances" may be accomplished in experiments, where one actually tries to execute all the phonological and phonetic processes by logical computations and computer-controlled hard- and/or software simulation of the physical sound production process (Holmes et al. 1964, Mattingly 1966, Rabiner 1968, Allen 1968, Lee 1968, Matsui et al. 1968, Teranishi and Umeda 1968, Coker 1968, Flanagan et al. 1970).

The principle of generating time functions for individual constituent features will involve the notions of target, coarticulation, undershoot, reduction, etc., in terms of articulatory states and movements, probably supplemented by a matching to pattern norms at the acoustic level. The overall process may be roughly characterized as follows:

(1) A table of inherent (target values for individual phonological units is given.

(2) Specification of complex (i. e. multi-dimensional) sequences of phonological segments is given for the linguistic form to be uttered.

(3) According to the partially universal and partially language-dependent system of phonological rules, logical interpretations of values of phonological units for individual feature-segment cells in the matrix are performed.

(4) Partially overlapping this logical process, physical interpretations are given by partially universal and partially language-dependent, as well as idiosyncratic rules for phonetic realization. This process is affected by random, as well as systematic, extra-linguistic variations.

The interpretive process will be described more concretely in reference to segments of various sizes, as well as phonological boundary symbols of various kinds. Inherent values may be specified for not only stationary target values, but also for transitional characteristics. Information that is utilized in the interpretive process generally is divided into a few categories: first, that inherent to the lexicon, perhaps including probabilistic characteristics of words; second, that for specifying syntactic (surface) structure, including grammatical formatives; third, surface-proper characterizations which are often suprasententially context-dependent, such as emphasis and contrast of different kinds, focus-presupposition status, etc. (Chomsky 1970, Takahasi et al. to be published). In this connection a synthesis-by-rule experiment employs novel principles for determing pitch values and durational and phonetic quantities (Coker and Umeda 1970, Umeda and Coker 1971). As for the prosodic features that are specified in the lexicon, accent-type specification and the related process of realization have been quantitatively discussed; Öhman (1968) has proposed an attractive

unified theory for Scandianavian dialects, and Fujisaki and Sudo (1971) made an elegant step further in modeling a transducer for realizing Japanese pitch contours.

The temporal structure of articulatory movements of speech organs is one of the central topics of recent speech research. Through studies of cineradiographic data, Öhman introduced a basic principle that goes beyond the traditional notion of segment sequence and smoothing and proposed that consonantal gestures be considered as an articulatory perturbation superimposed upon the basic time course for concatenated syllable nuclei (Öhman 1967). This notion seems to be justified by other cineradiographic observations of the tongue movements. Houde (Houde 1968a, b), in particular, suggested an account for the inherent vertical movement of the tongue for palatal/velar stops that may explain the so-called allophonic variations of the consonant in terms of the physically superimposed vowel gestures. We may speculate further that the consonantal gestures are given through particular muscular mechanisms that are separate from those for vowel gestures. Thus the raising of the tongue hump for [g] may be performed by an essentially different muscular mechanism than the apparently similar gesture for [u]. There are still some factors that complicate the actual trajectory of the tongue movements; further experimental studies will have to clarify this point of basic interest.

Coker and his coworkers simulated complex tongue movements in their synthesis experiments and also introduced an important notion in quantifying the coarticulatory processes. It is what they call "priority strategy" (Flanagan et al. 1970), through which an hierarchical treatment of articulatory features in respect to the timing of selected variables is implemented. Thus a labial stop, for example, is characterized by the feature of labial constriction, and this has to be realized, if necessary as a form of invasion into the adjacent (or even further) time segment for triggering the pertinent gesture. Two components with substantially different temporal characteristics of labial articulation are reported to be necessary--lip protrusion, in particular, for vowels and the semi-vowel /w/; and lip opening (constriction) primarily for consonants (Flanagan et al. 1970).

In short, our problem is to differentiate between linguistically controlled gestures and universal physiological and/or physical constraints. Our present knowledge is obviously insufficient to accomplish this partitioning. We may say, reviewing the works mentioned above, that we have some fragmental observations and even partial answers that narrow down the domain for future search.

The study of speech phenomena, as we have seen above, is also characterized by an inherently interdisciplinary approach. This is not surprising, because we must deal with the most intricate aspects of human mental activity as reflected in complex physical phenomena.

Summary

Problems we encounter in studies of acoustic properties of speech phenomena for understanding their relation to the linguistic code are discussed. A basic theoretical framework of the acoustic system of speech production is reviewed, and some of the research results on the vocal tract transfer characteristics are presented. Recently developed techniques of observing the mechano-physiological processes of speech production are described, and some tentative results concerning materials including Japanese and Korean are discussed in relation to some basic hypotheses in phonological theory. The role of synthesis by rule experiments in speech research is emphasized in connection with the problems of temporal organization of speech.

BIBLIOGRAPHY

J. Allen (1968), "Machine to Man Communication by Speech, Part II: Synthesis of Prosodic Features of Speech by Rule," Spring Joint Computer Conference, 1968, 339-344.

C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House (1961), "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Am. 33, 1725-1736.

T. Chiba and M. Kajiyama (1941), The Vowel, its Nature and Structure, Tokyo-Kaiseikan, Tokyo.

N. Chomsky (1957), Syntactic Structures, Mouton and Co., The Hague.

_____ and M. Halle (1968), The Sound Pattern of English, Harper & Row, Publishers, New York.

_____ (1970), "Deep Structure, Surface Structure, and Semantic Interpretation," Studies in General and Oriental Linguistics (R. Jakobson and S. Kawamoto, eds) 52-91, TEC Co., Ltd., Tokyo.

C. K. Chuang, S. Hiki, T. Sone, and T. Nimura (1971), "The Acoustical Features and Perceptual Cues of the Four Tones of Standard Colloquial Chinese," Proc. of the 7th International Congress on Acoustics 3, 297-300.

C. H. Coker and O. Fujimura (1966), "Model for Specification of the Vocal Tract Area Function," J. Acoust. Soc. Am. 40, 1271.

_____ (1968), "Speech Synthesis with a Parametric Articulatory Model," Preprints: Speech Symposium, Kyoto, A-4-1 - A-4-6.

_____ and N. Umeda (1970), "Acoustical Properties of Word Boundaries in English, J. Acoust. Soc. Am. 47, 94.

P. C. Delattre, A. M. Liberman, F. S. Cooper, and L. Gerstman (1952), "An Experimental Study of the Acoustic Determinants of Vowel Color: Observations on One- and Two-Formant Vowels Synthesized from Spectrographic Patterns," Word 8, 195-210.

_____, A. M. Liberman, and F. S. Cooper (1955), "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Am. 27, 769-773.

G. Fant (1956), "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," For Roman Jakobson, 109-120, Mouton and Co., The Hague.

_____ (1960), Acoustic Theory of Speech Production, Mouton and Co., The Hague.

_____ (1962), "Descriptive Analysis of the Acoustic Aspects of Speech," Logos 5, 3-17.

_____ and B. Sonesson (1964), "Speech at High Ambient Air-Pressure," Speech Transmission Laboratory Quarterly Progress and Status Report (Royal Institute of Technology, Stockholm) No. 2, 9-21.

_____ (1968), "Analysis and Synthesis of Speech Processes," Manual of Phonetics (B. Malmberg, ed.) 173-277, North-Holland Publishing Co., Amsterdam.

_____ (1971), "Distinctive Features and Phonetic Dimensions," Applications of Linguistics (G. E. Perren and J. L. M. Trim, eds.) 219-239, University Press, Cambridge.

E. Fischer-Jørgensen (1963), "Beobachtungen über den Zusammenhang zwischen Stimmhaftigkeit and Intraoralem Luftdruck," Z. für Phonetik, Sprachwissenschaft und Kommunikationsforschung Band 16, Heft 1-3, 19-36.

_____ (1968), "Voicing, Tenseness and Aspiration in Stop Consonants, with Special Reference to French and Danish," Annual Report of the Institute of Phonetics, University of Copenhagen 3, 63-114.

J. L. Flanagan (1965), Speech Analysis, Synthesis and Perception, Springer-Verlag, Berlin.

_____ and L. Cherry (1969), "Excitation of Vocal-Tract Synthesizers," J. Acoust. Soc. Am. 45, 764-769.

_____, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda (1970), "Synthetic Voices for Computers," Spectrum 7, No. 10, 22-45.

I. Fujii, O. Fujimura, and R. Kagaya (1971), "Dynamic Palatography by Use of a Computer and an Oscilloscope," Proc. of the 7th International Congress on Acoustics 3, 113-116.

O. Fujimura (1961a), "Motion-Picture Studies of Articulatory Movements," Quarterly Progress Report (Research Laboratory of Electronics, M. I. T.) No. 62, 197-202.

_____ (1961b), "Bilabial Stop and Nasal Consonants: A Motion Picture Study and Its Acoustical Implications," J. of Speech and Hearing Research 4, 232-247.

_____ (1961c), "Analysis of Nasalized Vowels," Quarterly Progress

Report (Research Laboratory of Electronics, M. I. T.) No. 62, 191-192.

_____ (1962), "Analysis of Nasal Consonants," J. Acoust. Soc. Am. 34, 1865-1875.

_____ (1967), "On the Second Spectral Peak of Front Vowels: A Perceptual Study of the Role of the Second and Third Formants," Language and Speech 10, Part 3, 181-193.

_____ (1967J), "Nihongo-no Onsē" ("Speech of Japanese"), Sōritsu 20-nen Kinenronbunshu (Collected Papers in Commemoration of the 20th Anniversary) (NHK, Radio and Television Culture Research Institute, ed.) 363-404 (in Japanese).

_____, H. Ishida, and S. Kiritani (1968), "Computer Controlled Dynamic Cineradiography," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 2, 6-10.

_____ (1970), "Current Issues in Experimental Phonetics," Studies in General and Oriental Linguistics (R. Jakobson and S. Kawamoto, eds.) 109-130, TEC Co., Ltd., Tokyo.

_____ (1971), "Remarks on Stop Consonants-Synthesis Experiments and Acoustic Cues," Form & Substance (L. L. Hammerich, R. Jakobson and E. Zwirner, eds.) 221-232, Akademisk Forlag, ISBN 87 500 1097 2.

_____ (1971J), "Hatsuon-no Katei" ("The Process of Utterance"), The Japan J. of Logopedics and Phoniatrics 12, No. 1, 11-21 (in Japanese).

_____ and J. Lindqvist (1971), "Sweep-Tone Measurements of Vocal-Tract Characteristics," J. Acoust. Soc. Am. 49, 541-588.

_____ and M. Sawashima (1971), "Consonant Sequences and Laryngeal Control," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 5, 1-6.

_____, I. Fujii, and R. Kagaya (1972), "Computational Processing of Palatographic Patterns," 1972 International Conference on Speech Communication and Processing, Boston, April 24-26.

_____, S. Kiritani, and H. Ishida (to be published), "Computer Controlled Radiography for Observation of Movements of Articulatory and Other Human Organs."

H. Fujisaki and H. Sudo (1971), "Synthesis by Rule of Prosodic Features of Connected Japanese," Proc. of the 7th International Congress on Acoustics 3, 133-136.

E. Gårding, O. Fujimura, and H. Hirose (1970), "Laryngeal Control of Swedish Word Tone - A Preliminary Report on an EMG Study -," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 4, 45-54.

M. Halle and K. N. Stevens (1962), "Speech Recognition: A Model and a Program for Research," IRE Transactions of the Professional Group on Information Theory IT-8, No. 2, 155-159.

_____ and K. N. Stevens (1967), "On the Mechanism of Glottal Vibration for Vowels and Consonants," Quarterly Progress Report (Research Laboratory of Electronics, M. I. T.) No. 85, 267-271.

_____ and K. N. Stevens (1971), "A Note on Laryngeal Features," Quarterly Progress Report (Research Laboratory of Electronics, M. I. T.) No. 101, 198-213.

M. S. Han and R. S. Wietzman (1970), "Acoustic Features of Korean /P, T, K/, /p, t, k/ and /p^h, t^h, k^h/," Phonetica 22, 112-128.

W. Hardcastle (1969), "A System of Dynamic Palatography," Work in Progress (Department of Phonetics and Linguistics, Edinburgh University) No. 3, 47-52.

S. Hattori (1951J), Onseigaku (Phonetics), Iwanamishoten, Tokyo (in Japanese).

_____, K. Yamamoto, and O. Fujimura (1958), "Nasalization of Vowels in Relation to Nasals," J. Acoust. Soc. Am. 30, 267-274.

J. M. Heinz and K. N. Stevens (1965), "On the Relations between Lateral Cineradiographs, Area Functions, and Acoustic Spectra of Speech," 5e Congrès International d'Acoustique, A44.

H. Hirose (1971), "The Activity of the Adductor Laryngeal Muscles in Respect to Vowel Devoicing in Japanese," Phonetica 23, 156-170.

H. Hollien (1965), "Stroboscopic Laminagraphy of the Vocal Folds," Proc. of the 5th International Congress of Phonetic Sciences, 362-364.

J. N. Holmes, I. G. Mattingly, and J. N. Shearme (1964), "Speech Synthesis by Rule," Language and Speech 7, Part 3, 127-143.

R. A. Houde (1968a), "Perturbations in the Articulatory Motion of the Tongue Body," Reports of the 6th International Congress on Acoustics II, B-13 - B-16.

_____(1968b), A Study of Tongue Body Motion during Selected Speech Sounds (SCRL Monograph No. 2), Speech Communications Research Laboratory, Inc., Santa Barbara.

H. Ishida (1969), "An Audio-Digital Hybrid Magnetic Tape Transport," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 3, 67-68.

K. Ishizaka and M. Matsudaira (1968), "Analysis of the Vibration of the Vocal Cords," J. Acoust. Soc. Japan 24, 311-312.

R. Jakobson, G. Fant, and M. Halle (1951), Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates (Technical Report No. 13), Acoustics Laboratory, M. I. T.

R. Kagaya (1971), "Laryngeal Gestures in Korean Stop Consonants," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 5, 15-23.

R. Kent (1970), "A Cinefluorographic-Spectrographic Investigation of the Component Gestures in Lingual Articulation," Ph. D. Dissertation, Iowa University, also in News Letter (Department of Speech Pathology and Audiology, University of Iowa) 3.

C.-W. Kim (1965), "On the Autonomy of the Tensity Feature in Stop Classification (with Special Reference to Korean Stops)," Word 21, 339-359.

_____ (1970), "A Theory of Aspiration," Phonetica 21, 107-116.

S. Kiritani and O. Fujimura (1970), "A Preliminary Experiment of the Observation of the Hyoid Bone by Means of Digitally Controlled Dynamic Radiography," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 4, 1-7.

_____(1971), "X-Ray Monitoring of the Position of the Fiberscope by Means of Computer Controlled Radiography," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 5, 35-39.

V. A. Kozhevnikov and L. A. Chistovich (1965), "Rech: Artikulyatsiya i Vospriyatiye," Chapter II, Moscow-Leningrad.

P. Ladefoged (1967), Three Areas of Experimental Phonetics, Oxford University Press, London.

F. F. Lee (1968), "Machine-to-Man Communication by Speech, Part I: Generation of Segmental Phonemes from Text," Spring Joint Computer Conference, 1968, 333-338.

I. Lehiste (1970), Suprasegmentals, The M. I. T. Press, Cambridge, Mass.

B. Lindblom (1963), "Spectrographic Study of Vowel Reduction," J. Acoust. Soc. Am. 35, 1773-1781.

_____ (1965), "Jaw-Dependence of Labial Parameters and a Measure of Labialization." Speech Transmission Laboratory Quarterly Progress and Status Report (Royal Institute of Technology, Stockholm) No. 3, 12-15.

_____ and J. Sundberg (1971), "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement," J. Acoust. Soc. Am. 50, 1166-1179.

L. Lisker and A. S. Abramson (1964), "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," Word 20, 384-422.

J. F Lubker and P. J. Parris (1970), "Simultaneous Measurements of Intraoral Pressure, Force of Labial Contact, and Labial Electromyographic Activity during Production of the Stop Consonant Cognates /p/ and /b/," J. Acoust. Soc. Am. 47, 625-633.

S. Maeda (1971), "Conversion of Midsagittal Dimensions to Vocal Tract Area Function," 82nd Meeting of the Acoustical Society of America.

A Malécot (1968), "The Force of Articulation of American Stops and Fricatives as a Function of Position," Phonetica 18, 95-102.

E. Matsui, T. Suzuki, N. Umeda, and H. Omura (1968), "Synthesis of Fairy Tales Using an Analog Vocal Tract," Reports of the 6th International Congress on Acoustics II, B-159 - B-162.

I. G. Mattingly (1966), "Synthesis by Rule of Prosodic Features," Language and Speech 9, Part 1, 1-13.

P. Mermelstein (1967), "Determination of the Vocal-Tract Shape from Measured Formant Frequencies," J. Acoust. Soc. Am. 41, 1283-1294.

_____, S. Maeda, and O. Fujimura (1971), "Description of Tongue and Lip Movement in a Jaw-Based Coordinate System," J. Acoust. Soc. Am. 49, 104.

S. Öhman (1966), "Coarticulation in VCV Utterances: Spectrographic Measurements," J. Acoust. Soc. Am. 39, 151-168.

_____ (1967), "Numerical Model of Coarticulation," J. Acoust. Soc. Am. 41, 310-320.

_____, A. Mårtenson, R. Leanderson, and A. Persson (1967), "Crico-Thyroid and Vocalis Muscle Activity in the Production of Swedish Tonal Accents: A Pilot Study," Speech Transmission Laboratory Quarterly Progress and Status Report (Royal Institute of Technology, Stockholm) No. 2-3, 55-57.

_____ (1968), "A Model of Word and Sentence Intonation," Speech Transmission Laboratory Quarterly Progress and Status Report (Royal Institute of Technology, Stockholm) No. 2-3, 6-11.

J. P. Olive (1971), "Automatic Formant Tracking by a Newton-Raphson Technique," J. Acoust. Soc. Am. 50, 661-670.

A. P. Paul, A. S. House, and K. N. Stevens (1964), "Automatic Reduction of Vowel Spectra: An Analysis-by-Synthesis Method and Its Evaluation," J. Acoust. Soc. Am. 36, 303-308.

J. S. Perkell (1969), Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study (Research Monograph No. 53), The M. I. T. Press, Cambridge, Mass.

R. K. Potter, G. A. Kopp, and H. G. Green (1947), Visible Speech D. Van Nostrand Co., Inc., New York; later edition by Dover Publications Inc., New York.

L. Rabiner (1968), "Speech Synthesis by Rule: An Acoustic Domain Approach," The Bell System Technical Journal 47, No. 1, 17-37.

J. A. Rome (1964), "An Artificial Palate for Continuous Analysis of Speech," Quarterly Progress Report (Research Laboratory of Electronics, M. I. T.) No. 74, 190-191.

M. Sawashima (1968), "Movements of the Larynx in Articulation of Japanese Consonants," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 2, 11-20.

_____ and H. Hirose (1968), "New Laryngoscopic Technique by Use of Fiber Optics," J. Acoust. Soc. Am. 43, 168-169.

_____ (1970), "Glottal Adjustments for English Obstruents," Status Report on Speech Research (Haskins Laboratories, New Haven) SR-21/22, 187-200.

_____, A. S. Abramson, F. S. Cooper, and L. Lisker (1970), "Observing Laryngeal Adjustments during Running Speech by Use of a Fiberoptics System," Phonetica 22, 193-201.

_____ (1971), "Devoicing of Vowels," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 5, 7-13.

_____ and T. Ushijima (1971), "Use of the Fiberscope in Speech Research," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 5, 15-34.

_____, H. Hirose, T. Ushijima, and O. Fujimura (1971), "Devoicing of Vowels," Proc. of the 7th International Congress on Acoustics 3, 109-112.

R. W. Schafer and L. R. Rabiner (1970), "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Am. 47, 634-648.

M. R. Schroeder (1967), "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements," J. Acoust. Soc. Am. 41, 1002-1010.

S. Shibata (1968), "A Study of Dynamic Palatography," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 2, 28-36.

Z. Simada and H. Hirose (1970), "The Function of the Laryngeal Muscles in Respect to the Word Accent Distinction," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 4, 27-40.

_____ and H. Hirose (1971), "Physiological Correlates of Japanese Accent Patterns," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 5, 41-49.

M. M. Sondhi and B. Gopinath (1971), "Determination of Vocal-Tract Shape from Impulse Response at the Lips," J. Acoust. Soc. Am. 49, 1867-1873.

K. N. Stevens, S. Kasowski, and G. Fant (1953), "An Electrical Analog of the Vocal Tract," J. Acoust. Soc. Am. 25, 734-742.

_____ and A. S. House (1955), "Development of a Quantitative Description of Vowel Articulation," J. Acoust. Soc. Am. 27, 484-493.

_____ and A. S. House (1956), "Studies of Formant Transitions Using a Vocal Tract Analog," J. Acoust. Soc. Am. 28, 578-585.

_____(1960), "Toward a Model for Speech Recognition," J. Acoust. Soc. Am. 32, 47-55.

_____, A. S. House, and A. P. Paul (1966), "Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation," J. Acoust. Soc. Am. 40, 123-132.

_____(1971), "Linguistic Factors in Communications Engineering," Applications of Linguistics (G. E. Perren and J. L. M. Trim, eds.) 101-112, Cambridge University Press, IBSN 0 521 08088 6.

_____ (to be published), "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," Human Communication: A Unified View (E. E. David, Jr. and P. B. Denes, eds.).

J. Sundberg(1969), "Articulatory Differences between Spoken and Sung Vowels in Singers," Speech Transmission Laboratory Quarterly

Progress and Status Report (Royal Institute of Technology, Stockholm) No. 1, 33-42.

H. Takahasi, O. Fujimura, and H. Kameda (to be published), "Behavioral Characterization of Topicalization: What is Topicalization, from a Computer System Point of View," Proc. of the 1971 International Meeting on Computational Linguistics.

R. Teranishi and N. Umeda (1968), "Use of Pronouncing Dictionary in Speech Synthesis Experiments," Reports of the 6th International Congress on Acoustics II, B-155 - B-158.

H. Umeda and N. Umeda (1965), "Acoustical Features of Korean 'Forced' Consonants," J. of the Linguistic Society of Japan No. 48, 23-33 (text in Japanese).

N. Umeda and C. H. Coker (1971), "Some Prosodic Details of American English," J. Acoust. Soc. Am. 49, 123.

J. van den Berg (1958), "Myoelastic-Aerodynamic Theory of Voice Production," J. of Speech and Hearing Research 1, 227-244.