FUNDAMENTALS AND APPLICATIONS IN SPEECH PRODUCTION RESEARCH

OSAMU FUJIMURA
Murray Hill, New Jersey 07974 U.S.A.

AT&T Bell Laboratories

ABSTRACT

Current issues in speech production research are reviewed with some historical perspective. It is emphasized that recent progress in computational and experimental techniques has brought about a substantial change in the research methodology, and that the interaction between linguistic theory and the understanding of the nature of speech signals has substantially contributed to the progress in both abstract description and speech analysis and synthesis.

0. Introduction

In this paper, I would like to express my personal opinion about the direction of research in conjunction with a fairly wide variety of topics in speech production research. The point I would like to make is that we need a deep inquiry into the nature of speech, in its linguistic, psychological, physiological and physical aspects, taking full advantage of the emerging computational techniques, in order to pave the way for future industrial applications as well as to understand what speech is. I will argue that some basic concepts in the theory of phonology and phonetics must be revisited (*cf.* Fujimura [1980]).

1. Physical Process

1.1. Acoustical Theory

According to the acoustical theory of speech production [Fant 1960], the physical process of speech production comprises two basic components: (1) source signal generation: the process of producing the source airflow through the glottis typically for sonorant parts of the signal, and pressure variation anywhere along the vocal tract near constrictions for some of consonantal parts; and (2) vocal tract filtering: the linear process of converting the airflow/pressure source signals into outcoming acoustic waves that represent speech signals.

There have been challenges to the source-filter theory, claiming that the plane-wave assumption is not valid in reality when we consider the three-dimensional turbulence formation above the glottis [Teager 1983]. There is at least one experimental attempt at measuring the three-dimensional distribution of acoustic pressure within the vocal tract in vowel production [Firth 1986]. The Fantian acoustical theory is the only workable (approximation) theory available at present, however.

In particular, it is well known that the vocal tract transfer functions for different vowel articulation gestures can be effectively represented by the F-patterns [Fant 1956]. We have verified this using an acoustical measurement on normal subjects (see Fig. 1). This acoustic measurement of the natural vocal tract does not involve any dc airflow. To the extent the observed transfer characteristics compare with predicted characteristics of naturally produced vowel sounds, our theory captures the essence of the acoustic process.

Perceptually, also, it has been our experience for a long time that a series-type formant synthesizer captures all vowel characteristics in terms of the phonetic values that are familiar to us.

The role of formant transitions, associated characteristically with consonantal gestures, was convincingly demonstrated by researchers at Haskins laboratories [Cooper *et al.* 1952; Liberman *et al.* 1954; Delattre *et al.* 1955]. These and related subsequent experiments led us to believe that the quasi-static formant theory was effective enough to capture basic characteristics of speech signals. For further progress, however, I believe this point has to be revisited. There is enough evidence to suspect that our current signal processing technology that is commonly used in automatic recognition schemes, for example, does not capture some crucial information including very rapid transitions for consonantal identification. Also, our knowledge of inherent signal properties of occlusive consonants (see Blumstein & Stevens [1979]) has not been utilized sufficiently in such applications.

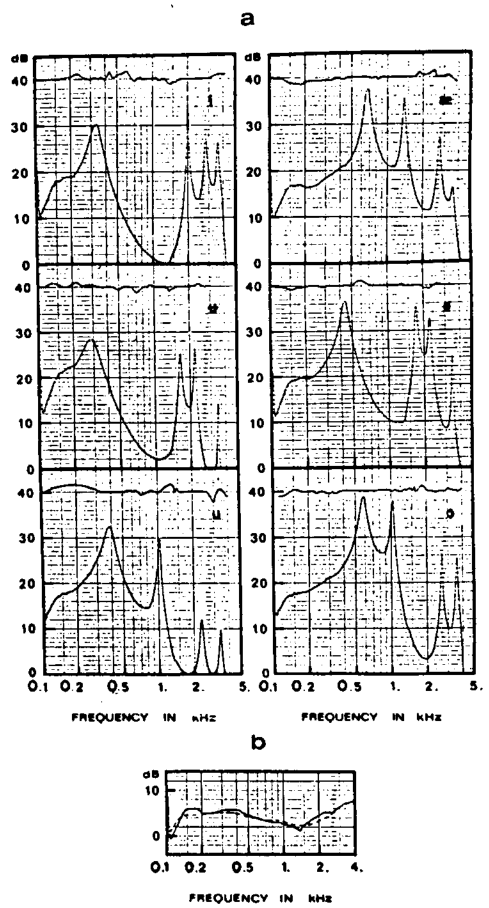

Fig. 1: Vocal tract transfer functions for Swedish vowels, estimated by a sweeptone method using a female native speaker. The curves include a constant frequency function (b), which is assumed to represent the transfer characteristic between the acoustic source (vibrator) output and the virtual excitation source above the glottis. The curve in the upper end of each frame represents the difference between the measurement and the theoretical prediction based on the series formant theory as in Fant [1961]. From Fujimura & Lindqvist [1971].

The synthesis experiments, incidentally, not only contributed substantially to our understanding of the nature of speech signals and their phonetic perception, but also, in combination with the subsequent emergence of the idea of analysis-by-synthesis [Stevens 1960], set a rather widely applicable methodology for studying complex human information processing.

1.2. Articulatory System

The mandible is literally a basic component of the articulatory system, and our understanding of its function in speech is still far from satisfactory. Edwards [1985] made a fundamental contribution to our knowledge in this field, clarifying its movement patterns in speech involving both rotation and translation relative to the skull.

The velum is presumably the simplest case for studying the phonetic functions relative to its physiological control and the resultant physical configuration. We seem to find a one-to-one correspondence between nasality as a phonological feature and the articulatory gesture of the velum, which probably can be effectively represented by a one-dimensional measure of velum lowering. The acoustical consequences of its movement, however, is by no means simple, nor is it limited to the coupling of the nasal tract to the (proper) vocal tract, as assumed in early works [Maeda 1983] (cf. House & Stevens [1956], Hattori *et al.* [1958]). Also, velum height is affected observably by the raised tongue dorsum for palatal consonants.

Because of the relative simplicity, the lip movement patterns have been the subject of quantitative studies by many investigators (see, for example, Bell-Berti & Harris [1982]). One particularly interesting topic is the relation of lip gestures to the mandible gestures. Macchi [1985] studied this problem in relation to segmental vs. suprasegmental functions using a statistical analysis of microbeam data. She found evidence that while both articulators contribute to the lip closure, some suprasegmental functions are related more closely to the mandible gesture than to the lip proper gesture.

The tongue is the most important articulator in the sense that it determines the largest portion of the vocal tract shape, with a large number of degrees of freedom, resulting in direct acoustical consequences. It is the most complex articulatory organ anatomically, physiologically, and physically. Its phonological implications are also complex.

Since the introduction of x-ray techniques, the laterally viewed midsagittal surface shape has predominated in the discussions of vocal tract modeling. The cylinder model of the tongue for speech synthesis [Coker 1968; Mermelstein 1973] appears to be the most extensively used computational algorithm for deriving area functions out of specifications of articulatory variables. I think it is now clear, however, that we cannot capture some of the most basic principles of articulatory control unless we consider the three-dimensional nature of the articulatory structures more directly.

Let me illustrate this argument with one example, just to demonstrate the nature of the problem [Fujimura & Kakita 1979]. In articulating a high front vowel, say [i], the tongue as a whole is pushed forward by the contraction of the posterior part of the genioglossus muscle. At the same time, some other muscles, including the contraction of the anterior part of the genioglossus, which run vertically near the midsagittal plane only, are used to form a fairly stiff surface shape with a significant groove along the midsagittal line. The resultant tongue surface is bulged upward on the sides. When the tongue is pushed upward and forward keeping this local condition, the sides touch the hard palate and support the tongue against a further upward forward push, leaving the central groove that forms a long and narrow open channel. The back of the tongue is forwarded considerably, creating a wide cavity behind in the lower and middle pharyngeal region. These conditions seem to be crucial for this vowel. The contact on the sides and the relatively rigid surface shape makes the articulation stable, without requiring excessive accuracy of the muscle contractions for forming such a critical narrow passage. This may be considered a viewpoint generalizing Stevens' [1972] concept of the quantal nature of speech production.

This study of the three-dimensional effects of muscular contractions has been performed by the use of the finite-element method of computational simulation [Kiritani *et al.* 1976; Kakita *et al.* 1985], and the interpretation above was inferred from this quantitative simulation work.

This tongue shape formation is inherently a three-dimensional process. It cannot be understood by considering the midsagittal configuration only, even though, after considering all these factors, we may well be able to compute the three dimensional shape, and thereby the area function of the vocal tract, accurately enough. This computation could be carried out, once we understand the mechanism, even from, say, positions of three appropriately chosen sample flesh points of the tongue surface in the midsagittal plane.

Interaction among different anatomical/physiological components of the articulatory system is a particularly difficult issue to study with limited available data. Intricate and often annoying effects of coupling between physical correlates of different linguistic variables have been observed. The segmental effects (of vowel identity, consonantal voicing, etc.) on voice fundamental frequency have been known for a long time (see Kohler [1986] for some relevant discussion). The tongue-larynx interaction has been discussed by Honda [1983], using careful electromyographic evaluation of activities of many relevant muscles.

The hyoid bone is located in between the tongue body and the larynx, conneted to both structures as well as the mandible via muscles and a unique sliding tendon mechanism. An interesting and powerful assumption, if it were true, were that this bone behaved as an effective positional stabilizer via various sensory mechanisms. A recent study by Westbury [personal communication], however, demonstrates, using a cineradiographic observation, that this assumption is not true. Rossi and Autesserre [1981] studied related issues concerning the intrinsic pitch of vowels, and provided a realistic picture of the interaction between laryngeal control and tongue gestures based on careful observations.

1.3. Source Signal Generation

Much research has been devoted to and progress has been achieved in understanding the mechanism of voice production. Publications are available, in part in proceedings of the Voice

Foundation Series of the Vocal Fold Physiology Conferences [Stevens and Hirano 1981; Bless and Abbs 1983; Titze and Scherer 1983; Baer, Sasaki and Harris 1987; Fujimura to appear]. The topics range from subcortical neural patterns (in animal vocalization) to computational modeling of the vibration mechanism. Notably, the anatomy, physiology and biophysics of the vocal folds themselves are substantially better understood in comparison to our knowledge, say, ten years ago, demonstrating the benefit of international and interdisciplinary cooperation.

The mucous membrane, the "cover" in Hirano's [1977] terminology, moves relatively independently from the muscular "body" of the folds, in the tangential direction in the speech mode, showing wave propagation along its vertical surface. Fleshpoints on its surface draw roughly elliptic trajectories. This two-dimensional picture (within the coronal plane) of vibration is not new in essense: Kirikae [1943] in his early study using a stroboscopic technique and carbon particles placed on fleshpoints of the muscosa observed clear wave propagation patterns from above in a living subject's larynx. Saito and his group at Keio University recentlly studied the movement patterns of fleshpoints in the cover and the body by special x-ray techniques applied to excised larynges [Saito *et al.* 1981].

Van den Berg [1957] originally discussed his experimental results about the interaction of the vocal folds with the airflow through them, revealing the basic physical principle of their vibration. Flanagan originated computational simulation of such a vibratory process using a mass and a spring to represent the vocal fold in interaction with the airflow [Flanagan & Landgraf 1968]. Ishizaka and Matsudaira [1968] theoretically demonstrated that realistic vibratory conditions under vocal tract loading can be explained only by adding another degree of freedom, and proposed a now classical two-mass model of the vocal fold vibration mechanism. This minimally approximates the three-dimensional structure by two pairs of mass and spring coupled with each other and with airflow. Flanagan and Ishizaka [1976] then produced a computational simulation of this model coupled with the vocal tract, demonstrating significant segmental effects of the vocal tract loading on the voice fundamental frequency.

Titze and Talkin [1979] approached this issue by using a more detailed part-by-part approximation. Fujimura [1981a] discussed the tension control mechanism based on the body-cover theory. Kakita *et al.*[1981] contributed biomechanical measurements of elastic properties of tissues. Titze and coworkers discussed various aspects of the vocal fold vibration mechanism such as energy exchange between the air and tissues [Titze 1985] and contributions of extralaryngeal factors to the voice fundamental frequency [Titze & Durham 1987]. Conrad [1987] proposed a functional interpretation of the fold vibration based on negative resistance as the key concept. Stevens [1974, 1975, to appear] also used circuit analogy and discussed large-amplitude as well as small-amplitude oscillatory characteristics. Rothenberg [1981] has contributed further insights into the interaction between the vibration and acoustic loading.

Fant [1983a] has been studying functional models of the source (volume flow) waveform in voicing using a unique system of parametric specifications, capturing important characteristics of the voice quality.

Experimental and computational studies of the vortices just above the glottis also are interesting from many points of view. It is a difficult area of experimental studies because of the small dimensions involved. Unfortunately, scaled-up physical experiments do not provide us with straightforwardly interpretable results.

With respect to the turbulent noise generation for fricatives, reader are referred to recent PhD dissertations by Shadle[1985] and Thomas[1985].

2. Physiological and Psychological Studies

2.1. Principles of Coordination and Control

One basic question in speech production research is what principle prescribes the time course of utterance, or the temporal pattern of motor commands for it, given the informational content of the message to be carried by it. The tacit assumption always has been that speech is a common daily activity for human life, and an uttererance must be economical in some sense [Lindblom 1983; Kent 1983]. Nelson[1983] faced this issue straightforwardly, and proposed a control-theoretical account. Based on this principle, combined with the concept of the quantal nature of speech production [Stevens 1972], Perkell and Nelson [1982] discussed some related stability issues of vowel articulation using microbeam data.

Among many topics concerning coordination of different organs in articulatory gestures, the concept of motor equivalence [Hughes & Abbs 1976; Abbs 1979] poses an interesting question with respect to high level planning and control in speech production [Abbs & Gracco 1982, 1983]. For example, a bilabial stop consonant inherently requires that the lips be closed, as its positional target gesture. For this condition to be (nearly) achieved, in terms of motor control, various patterns of activities of multiple muscles could be used. If for example the lip constriction gesture and the mandible raising gesture can be mixed in different proportions for the same goal, say lip closure, the proportional contributions of different articulators may vary from occasion to occasion. The question is what factors determine the variation and how we can describe the regularity involved.

Abbs and Gracco [to appear] report that in a repetition of a word 'sapple', the excursions of the upper and lower lips (measured at their vermilion borders) and of the mandible, for the vowel to consonant movement in the first syllable, are individually more variable than the resultant distance between the fleshpoints of the lips, over repeated utterances for each subject. They argue that such relative invariance of physical quantities that are directly related to the acoustic and perceptual consequences suggests that there is a strong role played by high level motor planning and adaptive (real time) control that combine the uses of different organs to achieve the given target gesture. According to them, the temporal coordination of a series of such target events also is fixed and is controlled at a level higher than for the movements of individual organs.

This is an appealing hypothesis. There are different feedback paths available for speech production, and they are often crucial for understanding aspects of normal speech. Unless we understand the way abstract planning and control are related to signal level phenomena, we may not be able to interpret the meaning of observed signals at an intermediate level, such as mandible movement as opposed to gestures inherently related to the control of lips proper. We may then ask which level of observation in the hierarchy of speech production control is most directly relevant to the description of the phonetic process. Random variation must exist, but identifying its existence is hardly sufficient, particularly when an independent mechanism such as saturation due to the direct contact of lips with each other contributes to statistical reduction of variability of the measured dimension. Macchi's work [ibid.], on the other hand, does suggest that variabilities of different component organs do reflect specific shares of different linguistic functions.

Several investigators recently proposed hypothetical principles for speech production coordination. Particularly at issue is how the temporal organization is designed, and what quantities remain invariant, given a phonetic identity of the speech material, resisting various causes of variation of the signals (see Perkell & Klatt [1986] for a collection of relevant discussions). Since the pioneering work by Lindblom [1964] and Öhman [1967], the basic concept of temporal organization for most investigators remained a concatenated series of positional target gestures representing phonemic segments, supplemented by a smoothiing process called coarticulation. In addition to this basic point of view, Kozhevnikov and Chistovic [1965] introduced a sequential planning model based on a statistical analysis of motor execution variation, and proposed a CV-type syllabic organizational unit. Henke[1966]'s look ahead model generalized the notion of coarticulation to include anticipation. In this connection, recently, Sternberg and his coworkers [1978, 1980] contributed a rather intriguing discovery about how the motor program for what appears to represent a stress group or a foot in English is formed prior to the utterance.

Kelso and his coworkers [1986] hypothesized a general speech production principle in accordance with a popular theory of neuromechanical control of biological systems [Haken 1977]. The basic idea is to assume simple oscillation as an underlying mechanism of speech production, and seek invariance in the phase relations among the underlying oscillatory movements of different articulators which form a task-oriented coordinative structure. They have conducted sets of experiments measuring relative contributions of lip-mandible gestures to bilabial closures. They go farther and argue that their result suggests some support of the consonant-vowel configuration as the basic phonological unit. (for my criticism and authors' reply, see [Fujimura 1986b; Kelso *et al.* 1986a].

When we consider apparent variance and invariance of specially designed and somewhat artificial (repetitive or purturbed) tasks, we need to be careful in interpreting data in different experimental situations. Different feedback paths may be used in different mixtures depending on the particular task and situation. Eliminating crucial dependence on one mechanism in one experimental situation does not lead us to conclude the lack of the use of that mechanism in other situations even for the same phonetic gesture. For example, repetitive utterance materials may introduce apparent characterization of movement control which may not properly belong to the nature of speech in general.

On the other hand, it is highly desirable, from a data-interpretation point of view, to design systematically controlled speech material, even at some cost of undetermined influence of the artificial contexts. A word paradigm, for example, comparing different vowel contexts for the same consonant in the same phonological environment, is never perfectly uniform with respect to, say, word familiarity, even if all words are natural existing words. I think we need to use both situations in such a case: natural linguistic materials in which items are not completely comparable, and systematically distributed artificial paradigms which must resort to some "phonetic performance" even by nonphoneticians, for the purpose of mutual calibration.

2.2. Neural Control of the Larynx and Sensory Mechanisms

Direct electrical access to higher level neural activities is not achievable in normal circumstances, at present, in spite of some promising new technologies such as highly sensitive magnetic field measurements. As for control of vocalization, however, animal experiments have made solid progress in our understanding about neural paths and control functions, for example relating activities at the brain stem level to laryngeal and other control in the monkey [Zealear 1987]. Sensory characteristics also have been studied by direct access to the afferent nerves. Davis and Nail [to appear] report on activities of both tonic and silent myelinated fibers of the internal laryngeal nerve of the cat, in response to carefully servo-controlled mechanical stimuli as well as chemical stimuli.

2.3. Observations in Pathologies and Speech Errors

One informative approach toward inaccessible human processes is to observe different types of pathological cases and compare them with normal cases. This is a rich and rewarding field, and in connection with the new research center with the microbeam facility at the University of Wisconsin, we expect substantial progress. For example, using the microbeam system at the University of Tokyo, Hirose and Kiritani [1985] obtained revealing data in cases of ataxia.

Another large area of study is speech errors. Recent studies take note of the fact that phoneme or feature value confusions between segments do not occur indiscriminately with respect to their role in syllable or word composition, and provide new framework for the description of the cognitive phenomena responsible for phonological performance [Kupin 1979; MacNeilage 1985]. Along the same line, developmental observations of child language and speech contain unique and valuable data.

3. Instrumental Methods

Algorithms of speech signal processing have become commonly available and several commercial systems exist for routine interactive studies of speech (acoustic) signals, using personal computers and workstations extensively. Major research groups often have more specialized advanced systems for efficient measurements of massive data. At the same time, large amounts of systematically collected speech materials are becoming available, with a large-scale effort invested into phonetic as well as some partial syntactic transcriptions of large databases such as the Brown Corpus [Frances & Kucera 1982] and the TI speech data base [Fisher *et al.* 1986].

3.1. Mechanical Measurements

The use of servomechanical adjustment of output impedance under flexible computer control for positional measurements of either flesh points or peripheral structures of articulatory organs provides us with a very powerful means for studying motor control mechanisms in speech [Muller *et al.* 1977]. For understanding the overall feedback functions under well-controlled mechanical conditions, such advanced techniques provide us with new possibilities of, for example, perturbation experiments, extending earlier explorations using the bite block conditions (see for example, Lindblom *et al.* [1979]).
Light-weight mechanical devices have been used by some investigators for obtaining the articulatory degreee of nasalization [Horiguchi & Bell-Berti 1984].

3.2. Ultrasonic Measurements

Surface contour information about organs that are not externally accessible can be obtained using ultrasonic techniques, which have seen good progress for clinical purposes. Ultrasonic pulse echo as well as penetration/nonpentration information (using the reflection of beams at the tissue-air boundary) gives us relatively good quality two-dimensional observations for some organs such as the tongue and the larynx, without causing any hazard such as ionization in the subject's body. Some investigators advocate the usefulness of ultrasonic measurements for detecting muscle contraction patterns as well as the surface shape of the tongue [Sonies *et al.* 1981]. As a novel application, Kaneko *et al.* [1981] observed minute vibration of the vo:al fold surface in response to external excitation.

The main limitation, in my opinion, of the ultrasonic technique applied to tongue observations lies in the mechanical loading effects on the outside skin. Ultrasonic signals are easily reflected at a boundary between a solid object or liquid and the air, and this necessitates a direct contact of the solid transducer surface onto either the skin itself or some liquid-like material as a transmission medium. This is particularly problematic for measuring movements, because of the inertia of such a medium, while it is circumventable for a carefully designed static measurement. The under surface below the floor of the tongue is quite soft, but it easily transmits force through the tongue, causing unknown dynamic deformation. With careful application, particularly in combination with other methods like x-ray microbeam for calibration, there is a good possibility of extensive use, however, since it can give different information related to the continuous surface contour as opposed to flesh-point sample positions of the tongue.

3.3. Optical Measurements

The use of a special fiberscope for laryngeal observations during speech utterances brought us new opportunities to understand the laryngeal gestures under phonetic control [Sawashima & Hirose 1968; Sawashima 1976]. Recently, in addition to the film and video recording methods in use in the past, a new technique of using a two-dimensional array structure of light-sensitive semiconductor elements (image sensors) has become feasible for high speed recording at a few thousands frames per second [Honda *et al.* 1985; Kiritani *et al.* to appear]. This makes it possible to digitally record glottal images without resorting to stroboscopic methods, which by definition is not very useful for studying any aperiodic characteristics of the vocal fold vibration.

3.4. General X-Ray Techniques

X-rays used to be the only source of information about dynamic tongue gestures, apart from the qualitative information obtained through visual inspection from the outside and subjective tactile and proprioceptive sensations. Because of the hazardous ionization effects in the body, however, the film method using fluoroscopic cineradiography (or other variants) is not generally recommended for extensive data collection of articulatory gestures. It also requires excessive analysis effort frame by frame. The video recording technique is probably significantly better but the situation is not qualitatively different. The basic problem stems from the flood exposure covering the entire image field. It should be mentioned, however, that careful and thorough examinations of limited amounts of film records in earlier years provided us with invaluable understanding of the physics and physiology of articulation [Chiba & Kajiyama 1941; Houde 1967; Perkell 1969; Wood 1979]. Some information with respect to the configuration in the lower pharyngeal regions, for example in relation to pharyngealization in Arabic languages, is also indispensable, at present, even though the available data are extremely limited. El Halees [personal communication] for this purpose used the recently developed xeroradiographic method, producing x-ray pictures of detailed structures with otherwise unimaginable clarity, but the extremely high dosage makes this method hardly applicable to more systematic studies. Rossi & Autesserre [1981] also applied this technique effectively for studying the functions of the hyoid bone.

The computed tomography [Kiritani *et al.* 1977] also provides invaluable information at the cost of a very high dose. It is possible, however, to reduce the required dose substantially, by readjusting the source intensity to barely sufficient amounts for distinguishing air from tissues, rather than using the normal conditions set optimally to differentiate tissue compositions. Another serious limitation of this method for speech research purposes is that the measurement time is inevitably very long, making even stationary vowel gestures somewhat difficult. In this respect, the nuclear magnetic resonance method has the same limitation.

3.5. X-Ray Microbeam System

Unlike the conventional film method, where flood x-rays emerge in a wide solid angle uniformly from a small x-ray generating spot on the target, the x-ray microbeam system uses a deflectable pencil beam of x-rays which is adaptively controlled by a digital computer. I invented the x-ray microbeam method out of the need to study dynamic articulatory gestures with the very minimum use of radiation and for practical feasibility of analysing extensive data. The first generation, a pilot system for testing the method, was implemented in 1968 at the University of Tokyo, with a 50-kV acceleration and a PDP-9 computer for control [Fujimura, Kiritani & Ishida 1973] (supported in part by NIH, USA). A second-generation device was implemented in 1973, with a 150-kV acceleration and a 2-mA electron beam current [Kiritani *et al.* 1975](Japnese governmental grant). This system was used for many data collection experiments, mainly by the University of Tokyo group, myself, and the speech physiology group at Haskins Laboratories in cooperation with the University of Tokyo group.

The third generation has been implemented at the University of Wisconsin, Madison, as the central research tool for a nationally shared speech research facility with research grants given by NIH (PI's: Abbs, Thompson and Fujimura, see Nadler *et al.* [1987]). This new system is designed for a 600-kV/5-mA operation, and is now being operated at 450-kV/5mA.

The reason for the high voltage is primarily twofold: (1) the geometrical design for distortionless image field requires a newly introduced transmission-type x-ray generator, and (2) to be able to cope with extraneous metal objects in the mouth, such as dental fillings, so that the experimenters are not excessively constrained about the choice of subjects in a wide range of experiments including studies of pathologies. In addition, (3) the energy absorbed by the body (*i. e.* ionization effects) is considerably less for the same detected energy, due to the better penetration of high energy photons.

The system is equipped with provisions for simultaneous acoustic and electromyographic data acquisition, and extensive uses by external groups are being scheduled under the coordination of a Users' Committee (K. S. Harris, chair).

A number of metal pellets (gold sphere or cylinder, one to three mm in cross dimension) are placed on the tongue and other articulators, and a few reference pellets are similarly placed on fixed points on the head (for head movement calibration and compensation). Pellets are searched by the microbeam automatically one by one time-sequentially, based on the past positions and according to prescribed prediction and search algorithms. In the new system, the exposure time for each position (pixel) is 2.5 to 10 microseconds, being adaptively chosen, so there will be no excessive radiation after securing a sufficient amount of photon detection. The effective frame-rate varies from pellet to pellet according to the experimenter's specification, and the microbeam is stopped by overdeflection for any moment it is not necessary for pellet identification. The radiation doses in realistic situations using the microbeam scheme are extremely small in comparison with any other x-ray methods.

In addition to obvious reasons for dose reduction due to selective exposures in space and in time, there are more subtle and still important reasons. Because of the use of the thin beam, the scatter photons are created only along the narrow beam, as opposed to the flood x-ray situation where they are created all over the volume of the exposed object, contributing to the summed-over noise registration. This results in a significantly better signal to noise ratio, and for this reason, the equivalent image quality is substantially superior. This, combined with the inherently high detector sensitivity, means that for a given task, even the local x-ray intensity at the point of exposure can be made considerably smaller than in a comparable situation (pellet position identification) using film methods.

The actual accumulative dosage in a few data acquisition sessions has been empirically evaluated using the Tokyo system. Dosimetry film and TLD mosaic have been placed on both the entry side and the exit side of the head to reveal accurate spatial distributions of accumulative dose within the image field, for two sessions each containing approximately 10-minute worth net total exposure. The total dose for such a typical session would be less than the accumulative cosmic ray exposure for the person under normal circumstances. The peak dose rate (averaged over a very small volume along the direction of photons) is really what we should pay attention to in planning experiments, taking a conservative attitude. It was found to be about 10 mR at maximum within the image field for each 1-minute worth net exposure. This means that if we take the local peak dosage as an index for conservative precaution, an hour long net or continuous data acquisition would amount to a peak dose roughly comparable to one dental bitewing shot.

3.6. Magnetic Methods

While the radiation hazards are minimized by the use of the microbeam, it would be nice if we could perform comparable tasks without using ionizing photons at all. Sonoda's early attempt used a small permanent magnet attached on the tongue, its position being determined by externally located field detection coils [Sonoda & Kiritani 1976]. This system has the basic limitation of not being capable of tracking more than one sample point simultaneously. The use of an externally created ac field picked up by a small detector coil in the mouth circumvents this constraint [Oka 1980; Schoenle *et al.* 1983]. Each detector is a 4 x 4 x 2.5-mm coil wrapped around a ferrite core with a pair of thin wires for external connections, and it is glued on the tongue surface as in the case of the microbeam pellets. Perkell and Cohen [personal communication] recently succeeded in tracking one "pellet" on the tongue yielding an extensive set of data. A practical system using a large number of "pellets" simultaneously remains to be developed in order to replace the x-ray microbeam for general purposes of articulatory studies. The crucial dependence on the attached wires leading to the outside measurement system does constitute a limitation. Also, the metal pellets for the microbeam system can be substantially smaller. The magnetic method does have a distinct advantage, however, in not being constrained by metal objects in the mouth such as dental fillings and caps, in addition to the nonuse of ionizing rays.

3.7. Electrical Methods

3.7.1. Palatography

Computerized palatography, in my opinion, is a very useful device for both research and tutorial/clinical purposes. It makes the traditional palatography applicable to moving gestures, and at the same time, the data are now recorded in computer files directly. The idea of using multiple electrodes embedded on an artificial palate, to my knowledge was first tried in Stevens' group at MIT by Rome [1964], who represented the time course pattern using the spectrographic display scheme. This dynamic palatography was then computerized using oscillographic displays [Fujii *et al.* 1971]. We studied characteristics of Japanese apical consonants [Fujimura *et al.* 1973a] and Miyawaki [1972] studied their palatalization using this method. Eek [1973] also applied computerized dynamic palatography to studies of Estonian palatalized consonants, revealing an intriguing difference in the temporal characteristics of phonetic implementations among languages.

Schemes using the same basic principle, called electropalatography or dynamic palatometry, are in use by several groups for phonetic research [Hardcastle 1972, 1974, 1984; Fletcher *et al.* 1975; Sawashima & Kiritani 1985], and for clinical applications [Shibata *et al.* 1979]. The device is now commercially available with new features, particularly with the provision for using ready-made palates as opposed to the palate specially made for the individual.

3.7.2. Glottography

Electroglottography [Fant *et al.* 1966; Smith 1981; Childers *et al.* 1984] and laryngography [Fourcin and Abberton 1977] have been used extensively for phonetic studies of vocal fold vibration patterns. While it is only an indirect indication of the condition of the vocal fold contact, its fast response and the lack of invasive elements makes it practically useful for many situations where other more direct methods of observation are not applicable.

3.7.3. Electromyography

Measurements of the muscle activities are at present the best we can do for directly observing physiological patterns above the physical levels in speech behavior. The use of hook-wire electrodes prevails in electromyographic studies [Hirano & Ohala 1969]. The interpretation of the signals representing contributions from the complex of the muscle fibers under unidentified physical conditions is difficult for rigorous quantitative discussions (see for a careful and elaborate method of single motor 'unit decomposition, Deluca [1975]). With appropriate care, EMG is the most powerful means for assessing speech control principles via direct measurements (see Fujimura [1979] for a review of its applications in studies of laryngeal control gestures). Combinations with other methods of physical observations are often desirable, and the new research facility at the University of Wisconsin aims at simultaneous digital data recording with 'the microbeam pellet position measurement.

4. Temporal Organization and Linguistic Structure

The general aim in this area of study is to separate physical constraints from linguistically motivated control. My own approach around 1960, working in Halle and Stevens' group at MIT, was to observe the articulatory dynamics as much as possible, through high-speed motion picture recording and analysis of the lip movement [Fujimura 1961]. Öhman [1967] in the same line of effort, working with Lindblom [1968] at MIT and then KTH (RIT, Sweden), analysed x-ray data of tongue movement as well as acoustic data, and proposed a quantitative model formalizing a now standard concept of coarticulation. At the same time, Öhman proposed the perturbation theory of consonantal articulation, introducing an important conceptual deviation from the classical notion of speech as a single chain of segmental units. He tried to quantify the inherently multidimensional nature of speech, by a method which later would have been called a projection principle.

4.1. Segment Concatenation and Coarticulation

Coarticulation in the Lindblom and Öhman's sense is basically the process of parameter smoothing in the physical realization of a string of phonetic segments [Stevens 1983]. If we take the phoneme to be the segmental unit, however, and expect an observable speech signal or its conventional parametric representation (as in speech synthesis experiments) to be constructed by concatenating segmental target values into a string, it does not capture some important characteristics of natural speech. The concept of coarticulation as a smoothing filter for any parameter, such as formant frequency, quite possibly with some notable asynchrony allowed, can be generalized to include the more traditional and qualitative linguistic notion of assimilation, or what we might call soft coarticulation [Fujimura & Lovins 1978]. This makes the string concatenation model more tenable, but at the same time it makes it more difficult to assess its validity; and still, it is difficult to explain observed *ad hoc* variation of phonemes in different environments *[ibid]*.

An appropriate model of concatenation and smoothing, in my opinion, can be obtained only if we describe the production system using a temporal pattern comprising multiple dimensions, each of which is related to a physiologically controllable variable. The mapping relation between such a set of control variables into the conventional speech signal parameters such as formants and pitch is likely quite complex, involving nonmonotonicity and hysteresis. Also, the control program itself is under the influence of feedback and anticipation. We need to know what these mapping characteristics are, or at least what qualitative constraints they have, before we can determine what the effective variables are for successfully relating abstract linguistic units to physical phenomena. It is a horrendous task to pursue, but recent progress in technology, particularly in computational methods, has made us feel that some progress is in sight (see for examples of research efforts along this direction, Coker [1968]; Browman & Goldstein [1985]). It would not be possible at all, however, if we had to rely entirely on the inductive approach. Since the superpositional principle is not expected to work over the independent variables unless we find an effective transformation, resorting to statistical approaches blindly does not look very promising. Fortunately, recent progress in phonological theory as referred to later, gives us good insight into this issue, and of course, in turn, any discovery in the facts of speech will contribute substantiallly to the formulation of a successful theory of phonology.

In this connection, from an engineering point of view, I believe the optimal choice of a phonetic unit as long as we remain in the segment concatenation method, is the demisyllable or something equivalent [Fujimura 1976; Fujimura *et al.* 1977a; Browman 1980; Macchi 1980]. The demisyllable was also successfully adopted in automatic speech recognition [Rosenberg *et al.* 1983]. The basic reason for the efficacy of demisyllables is that the predominant types of context sensitivity of phonemes, *i. e.* many sorts and degrees of allophonic variation, some nopelessly *ad hoc*, are effectively contained within the domain of the demisyllable apart from the prosodic effects (see *infra*).

Another apparently similar technique is to use phoneme diphones [Sivertsen 1961; Dixon 1968; Olive 1980] as the "segmental" unit for speech synthesis. Being based on the phonemic theory, this approach is originally independent from the demisyllabic one, but in practice both techniques in speech synthesis have been converging using about the same number of units stored in the inventory. Olive's diphone approach has many additional features as well as elaboration in details.

4.2. Phonology and Phonetics -- Intonation and Other Topics

One important recent development in the theory of phonology that bears strong implications on understanding the temporal organization of speech is the trend toward integrating phonetic observations with the very core of the theoretical discussion. This new trend is most strongly seen in the description of intonation/accent patterns, but it can now be found in the entire domain of (nonlinear) phonology, influencing the basic structure of phonological representation from the lexical level down. Articulatory data, collected systematically with careful speech material designs guided by the basic theoretical interest of linguistic structures may soon constitute unique objects of such discussions.

The spirit of nonlinear phonology at least in the case of so-called suprasegmental description is nothing new (see *e. g.* Hattori 1961]), and there has been a relatively long tradition of descriptive work on intonation in Europe involving different experimental methods (see for more recent examples, Vaissiere [1977]; Nishinuma & Rossi [1981]; Gårding [1983]; Thorsen [1984]; for discussion of interacting factors see Nooteboom & Terken [1982]; see also Eek (ed.) [1978] for reports on various studies on different languages). For its theoretical impact in general and formal phonology, we had to wait for the most recent progress using advanced computational environments. Some of the full-scale experimental effortss on sentential intonation by those familiar with linguistic theoretical issues was triggered by astute observations by speech researchers with engineering backgrounds (see *e. g.* [Maeda 1976]). After Liberman [1975]'s theoretical lead (see also Liberman & Prince [1977]), Pierrehumbert's dissertation [1980] established a new experimental/computational methodology of phonological/phonetic studies.

Traditionally, according to the explicit formulation due to generative phonology [Chomsky & Halle 1968], phonological rules constituting a precise body of formal specifications dealing with discrete symbols (specifically binary-valued distinctive features) produced an input to the process of phonetic implementation, which handled numerical or continuously valued variables representing physical correlates of those features. The objects of the entire phonological manipulation were the feature matrices, which separated phonemic segments as "simultaneous bundles" [Jakobson *et al.* 1951] represented by its columns. At the output level of phonology, the so-called systematic phonetic representation used numerical specifications of feature values, as a buffer representation between the symbolic and numerical computations. I do not believe that such an independent level of description is tenable [Fujimura 1970; Keating 1985]). It is now an empirical question if the separation of numerical processing from symbolic manipulation as subcomponents of a body of ordered rules or processes can be maintained (see Ladd [1986] for relevant observations). It is conceivable that the two distinct subsets of rules are not separate in terms of rule ordering but different in their formal properties.

The concept of simultaneous bundles is abstract, just as that of distinctive opposition is. The current argument is that the representation at the most abstract (lexical) level has to be inherently mutidimensional, different features covering different (abstract) temporal domains, and dimensional structures must reflect some articulatory functions [Clements, personal communication; Halle 1983, 1985]. The emerging theory of melody-skeleton association seems to provide a good bridge between our findings about articulatory movement patterns and their temporal organization on the one hand, and the abstract phonological representation necessitated out of distributional and derivational observations on the other [McCarthy personal communication; Fujimura 1986c]. At the same time, phonological representations may specify linguistically significant oppositional values abstractly and sparsely, as opposed to completely segment by segment. Such a descriptive system like those based on a marking convention [Chomsky & Halle 1968; Kean 1975], for example, may make good sense particularly if it is assisted by the syllabic framework and conventions involving resyllabification, for example a scheme proposed by Borowsky [1986].

A point of dispute, given this relaxation of the one-dimensionality (*i. e.* concatenative linearity) or the "simultaneous bundle" constraint, and the introduction of any rather specific but complex structural framework, is whether the abstract feature specifications should be given unit by unit completely, or rather specifications are inherently nonsegmental in the sense that they (whether quasi-static values or dynamic patterns as the "target" configurations) are sparsely specified in the multidimensional space down to the level where numerical implementation rules operate. In the latter case, the realization rules would compute the entire time course of each dimensional variable to specify the temporal course of physical signals for a large phrasal unit. Pierrehumbert (see *infra)* clearly takes the latter view, whereas Inkelas and her coworkers [1987] maintain the former view discussing African tone/intonation phenomena.

A point of future study related to this topic is the nature of phrasing in speech utterance. A three-level framework of phrasing hierarchy has been proposed by Pierrehumbert and Beckman [in press] (also Beckman and Pierrehumbert [1986]) in their studies of. Japanese and English. In Japanese, within the minor (accent) phrase, any but the first lexically specified accent marks lose their realization, according to traditional accounts (see for rule formulation, McCawley [1968]; Haraguchi [1975]). This is usually interpreted as an erasure of such marks. Recently, a concept of catathesis was proposed by Poser [1984] (in conjunction with Pierrehumbert's descriptive framework) relating pitch contour realizations in contiguous phrases. This process, unlike the so-called pitch declination, is conditioned crucially by the existence of accent in the preceding phrase. When we handle a larger phrasal unit, according to the catathesis theory, a qualitatively similar phenomenon takes place, but the effect is not to eliminate the mark nor ignore it, but to reduce its manifestation for the subsequent phrases, if and only if there is a preceding accent (in the preceding minor phrase). This raises the following question: Is the accent deletion really a symbolic phonological operation, or is it only a relatively strong degree of reduction? Further, it could be questioned if the distinction between the smaller and larger phrasal units are something of a categorical nature, as expected from the syntactic motivation of the phrasal structures, or is it to be captured (roughly speaking at the phonetic level) as continuously varying boundary effects? That a complex set of discourse factors influence the boundary effects in numerical manners may favor the latter point of view, and as far as I know, there is no evidence contrary to this.

Another set of observations being discussed in terms of the relation between phonology and phonetics concerns the neutralization of phonemic distinctions in certain phonological environments. Dinnsen and Charles-Luce [1984] studying final obstruents in Catalan challenged the separation of phonology and phonetics, claiming that the phonological rule devoicing obstruents must apply after the phonetic implementation rule that accounts for speaker-dependent final devoicing. Similarly, the final consonantal tense/lax or voiced/voiceless opposition in German has been studied by several investigators, both in production and perception [Fourakis & Iverson 1984; Port & O'Dell 1985]. The concept of neutralization was revisited by Fourakis [1984].

Keating [1985] discussed the same difficulty in her study of vowel duration and voice onset timing patterns and has proposed a modification of the theoretical framework, allowing the grammar of a language to control "all aspects of phonetic form". This view may seem necessary to explain what is observed using the traditional framework of phonetic description. The question crucial to the theory of phonology is, however, not just what is sufficient for the description of the observed patterns, but how we can transform observable signal characteristics to units and structures that are effective for phonological representation. If we do not pursue an answer to this linguistic question, we will simply have to yield to the more complex data as we become capable of the more sophisticated measurements.

What is important here, however, is the fact that the fundamental concepts of phonological representation are being challenged, as the result of accurate enough quantitative observations of actual physical signals, together with the technical capability of comparing exactly realized complex mathematical schemes by computation. The conceptual process of speech synthesis by rule is a concrete technical experience in our present-day research environment, and it has emerged, in part, as the result of an engineering interest in a machine that relates lexical representations (often given in orthographic text) to speech signals.

4.3. Articulatory Aspects of Prosodic Control

Traditionally, prosodic effects on speech characteristics have been discussed in connection with their manifestation in voice pitch modulation and temporal patterning of segmental units. Thus, the segmental units (phonemes in most discussions) displayed their inherent physical correlates when they were concatenated into a temporal string, with the coarticulation or smoothing with the resultant reduction or undershooting as the only modification, while pitch and durational modulation were superimposed onto this representation of the speech signal. Some minor (presumably universal) interactions of laryngeal control with articulatory characteristics have also been considered. This picture is typically represented in the tradition of speech synthesis by rule [Liberman *et al.* 1959; Holmes *et al.* 1964].

Recent studies clearly show that this classical view only reflects a lack of precise enough data, or at best, careful avoidance by the phoneticians of the intruding complexity of "nonessential" factors in the phonetic description. Every phonetician has known that samples of vowels could not be collected from different contexts, segmental, suprasegmental, or extralinguistic, for physical measurements to yield valid comparison of contrasting phonemes. Even a narrow phonetic transcription cannot be performed mechanically, because supposedly identical phonetic segments of a language would vary from one condition to another. The engineering interest in designing a machine to identify words for practical purposes has compelled us to confront this outstanding problem (for some relevant discussions, see Fujimura [1984]; M. Ohala [1983]).

4.3.1. Focus and Phrasing

Contrastive emphasis placed on a particular word in a sentence utterance introduces remarkable effects not only in the pitch contour and segmental durations, but also in the articulatory movement patterns including what appears to be the target position. Fig. 2 illustrates an example of the vertical movement of a metal pellet placed on the blade of the tongue, tracked by the x-ray microbeam system at the University of Tokyo. The subject was a phoneticaly trained female speaker of a dialect (Georgia) of American English.

The two utterances demonstrate the effects of different placements of focus (contrastive emphasis). It can be readily seen that the affected words are uttered with radically different gestures. The syllable nucleus of the word "six" (see the single arrow) shows more than three times as deep a valley in utterance (a), accompanied by a considerably extended time interval between the downward and upward (transitional) movements. The

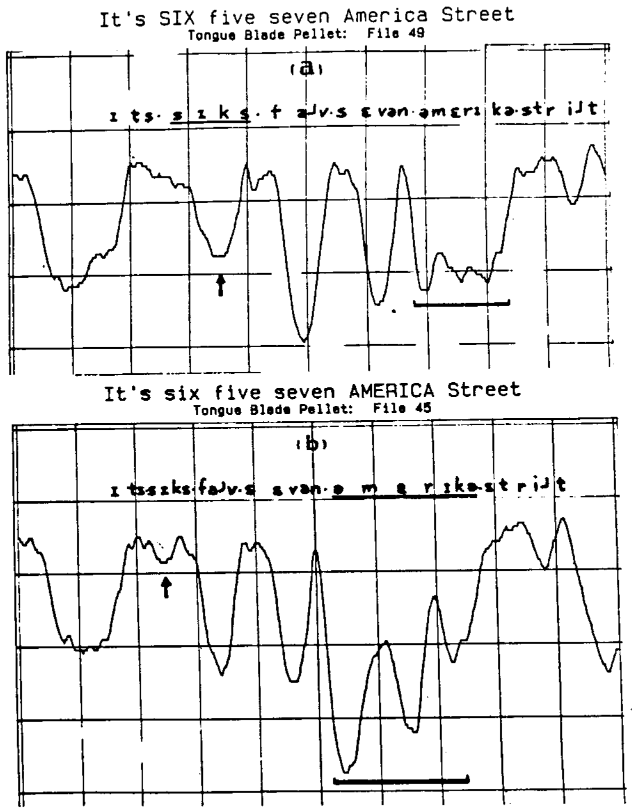

Fig. 2: Tongue blade movement (vertical position) as recorded by x-ray microbeam tracking. 'It's six five seven America Street' spoken by a female American speaker, with focus placed on (a) 'six' and (b) 'America'.

consonantal gestures on both sides of the valley also show some differences between the two versions. The portions of the utterances representing the word 'America' (see the bracket), on the other hand, demonstrates an even more dramatic difference in gesture. There, the observed tongue blade height patterns share almost nothing between the two versions. Finding correspondence between the two curves as shown by the transcription in the figure, is difficult without resorting to an examination of other pellet positions (and the acoustic signal), in spite of the fact that the two front vowels as well as the /r/ all presumably involve some inherent tongue blade gestures.

In this experiment, a few utterances were recorded for each of the four conditions placing emphasis on different focusable words. The patterns were observed to be qualitatively very consistent among different utterances for the same emphasis condition, but the gestures for emphasized words tended to vary considerably in terms of the extent of the excursion and temporal expansion.

A somewhat similar modulation with less change in the depth of nucleus valleys was observed when distinct phrasing patterns were used for the same word sequence in arithmetic formulae such as

(5 + 5) x 5 (yielding 50) vs. 5 + (5 x 5) (yielding 30).

4.3.2. Iceberg Patterns

In the course of studying temporal organizations of articulatory movement patterns, we realize it is rather difficult to define reliable land marks which we can rely on in comparing utterances of the same phonetic segmental material under different environments. The familiar notion of a segmental boundary (for representative examples see Lehiste [1970]; Umeda [1975]) displayed as acoustic events such as voice onset, consonantal implosion and explosion, do not find corresponding discontinuities in the articulatory time course. This is particularly true, presumably, because we use selected flesh points on the articulators such as the tongue blade, which, depending on the context as well as the particular phoneme, may or may not represent the point of articulation. For a precise timing definition of an event that is crucial in terms of the acoustic consequences, we will have to refer to a three-dimensional measurement covering a wide spatial domain, as seen in dynamic palatographic studies.

In my opinion, however, the apparent difficulty reflects a deeper issue. The dynamics of articulatory structures is inherrently continuous, involving a set of finite quantities like force and mass. Even when the velocity of a particular part of the organ changes abruptly, for example by collision with a heavy hard structure such as the palate, the central part of the organ keeps moving rather smoothly. Apart from the possible indirect reaction through neural feedback, what determines the time course of the entire system reflecting the neural commands is the physically central rather than peripheral part of the structure.

Also, for the purpose of quantitative analyses, a smoothly changing variable is mathematically more tractable than discontinuous functions, because smooth functions can be handled at least locally by a linear approximation. This means that within a selected range of change, the system can be treated as a superpositional system, where different factors can be easily separated out by controlling contributing factors one by one. In a formidably complex process such as speech production, this is perhaps the only practical initial approach, until we have some comprehensive view of the entire system with respect to interrelations among specific parts of the system.

One more reason favoring recording smoothly changing variables is that our measurements are always noisy. A discontinuous time derivative used as the means for evaluating the crucial event is inherently susceptible to errors due to small noise in position measurement. Especially if the purpose is to determine timing values of crucial events, continuously moving parts of the time functions provide the most accurate evaluation of timing, in comparison with, for example, an evaluation of the time when a movement starts from the standstill condition. If we define an event of invariably fast movement of a sample point fixed to the structure, say a pellet, crossing across a prespecified position, say height threshold, then the accuracy using positional measurement is very high with respect to the time evaluation. In order to make phonetically meaningful measurements, however, we have to find a crucial condition, say a specific value of height for the selected flesh point that makes sense as a definition of a phonetic event.

If we have a validated model of the time course for the given observable quantity, say if the system behavior is known to be determined by second order dynamics [Fujisaki 1977, 1983], then we can use a large segment of the time changing variable that covers an interval during which system parameters can be assumed to take constant values, for a semiglobal curve fitting procedure. This is a very noise-resistive method. Some recent temporal studies (*e.g.* Ostry *et al.* [1983]) in effect assume such a simple model (locally sinusoidal change).

My approach is to try to find relatively invariant movement patterns that can be operationally defined reliably enough for the purpose of timing evaluations of landmark events. This method was motivated by the informal observation of various data from microbeam measurements. For some parts (in terms of pellet height) of the movement of the crucial articulator for place-specified consonants (the lower lip for labials and the tongue blade for apicals), fairly reproducible results seemed to emerge with respect to timing modulation of such events as the result of prosodic control [Fujimura 1981, 1986]. Using a special statistical process to automatically and empirically decide such positional ranges for a selected domain of prosodic variabliity, we evaluated, for sets of data described above relative to focus and phrasing, the temporal modulation relation of each pair of utterances (see Fig. 5). The data are only preliminary, and await further verification using more data, which hopefully will become available very shortly from the Wisconsin microbeam system.

For such patterns that seem to be characteristic of the consonant-vowel combination, or more generally for a given demisyllable, where the observed articulator is crucial for the place specification of the consonant, I gave the name "iceberg", because such a movement pattern floats around fairly freely in time relative to other articulators' movement patterns, when segmental or prosodic contexts change [Fujimura 1981].

4.3.3. The Case of Velum Movement

Vaissiere [personal communication], using the microbeam data from the University of Tokyo, studied the velum movement patterns in utterances of several sentences, as well as words in isolation, spoken by two native speakers of English (General American). She interpreted the time functions representing the vertical position of a sample point of the velum surface, obtained by tracking a pellet attached on a flexible plastic strip which was placed on the velum in the nasal cavity [Fujimura, Miller & Kiritani 1977]. In prescribing the time course of velum height, she defined the "strength" of the oral consonant with respect to its tautosyllabic effects. The strength is conditioned by intrasyllabic position as well as stress. For the positional target, she concludes tentatively that there is no target values for vowels, varied positions being specified for both nasal and nonnasal consonants depending on nonsegmental conditions.

One particularly interesting observation she has made is that the strategy related to syllable reduction seems to vary basically from one speaker to another. In one speaker, the movement reduction for prosodically weak position seems to be explained by undershooting due to time constraints, while for the other speaker, velocity seems to be under control independently. It is hoped that such issues will be pursued with extensive data using many subjects in different languages.

In many languages, it has been reported that velum height for word initial position is higher than for word final position, segmentally (nearly) *ceteris paribus* [Ushijima *et al.* 1972; Fujimura 1977]. Some observation using my own articulation in Japanese shows that this initial vs. final distinction is observed for intrasyllabic position even when the nasal consonant is in word-medial position.

4.3.4. Allophonic Variation

One important issue that stems from the traditional segmental view of speech is the allophonic variation of the same phoneme depending on the context. Presumably, any universal effects of (hard) coarticulation are excluded from such descriptions of segmental variation, but that does not mean that the remaining aspects of coarticulatory processes do not involve utterance parameters. Parameters such as time constants of movement patterns, inherent strengths of influence over neighboring elements within an articultory dimension, susceptibility of a target position or a movement pattern to such influences, must vary from language to language, dialect to dialect, and part of it may well vary speaker to speaker. The patterns of use of particular articulators for the same phonological functions may also vary, as we have seen in Vaissiere's observation of velum movement patterns. Furthermore, parameters specifying a neutral (rest) position, range of movement, sensitivity to prosodic modulation, etc. of the articulators must be specified as to what we may call "phonetic disposition" to characterize each language, dialect, ideolect, etc. In order to compare different linguistic systems, in terms of phonological patterns implemented as speech, we need a complex and very sophisticated normalization method to be applied to different phonetic systems.

Precise descriptions of coarticulation and normalization processes are not known to us at present, but as a matter of principle, we may assume such well-defined processes which we could use to identify unexplained variation of phonetic values of phonological units, phonemes or syllables. A large part of such variation would be related to prosodic effects. There are known salient cases of phonetic variation of phonemic segments, however, which can be recognized as *ad hoc* in the sense that any language dependent assimilatory principle (*i. e.* even soft coarticulation) would not be expected to predict them[Fujimura & Lovins 1978]. I think most of such known allophonic variation is contained within the domain of the syllable, or in fact, the demisyllable.

My interest now is if we can find out some parts of such seemingly *ad hoc* variation to be describable in terms of a more general systematic (but of course language dependent) description of temporal characteristics of articulatory processes. I think the following observation of American English flapping may be suggestive of such a possibility.

In American English, intervocalic /t/ and /d/, typically in a stressed-reduced environment (as in 'better', see Kahn [1976]; Laferriere & Zue [1977]), are pronounced with a transient and incomplete closure accompanied by voicing for both /t/ and /d/ (tap or so-called flap, see Ladefoged [1977]). The microbeam observation with respect to the tongue blade pellet (about one cm behind the tip of the tongue) has revealed a very interesting dynamic characteristic of this articulatory gesture. Fig. 3 shows a comparison of a minimal contrast between a voiceless stop and a (voiced) tap for a pseudo-English phrase, spoken by a female speaker.

The two sets of time-functions representing coordinate values of pellets are aligned in time, in such a way that the two utterances show a fair agreement roughly, apart from the following two points [Birnbaum, personal communication]: (1) The mandible shows some raising for the stop gesture but not for tapping, and there is some tongue body movement for the stop correspondingly. (2) The tongue blade (presumably tip also) shows a distinctly different type of gesture both in the time function shape and the timing of the event as a whole relative to other articulators' temporal patterns.

Point one probably can be explained in terms of the difference in the use of the linguamandibular gesture related to both the phonological syllable margin status and the physical constraints or the physiological mechanism used for forming the apical closure. It should be emphasized that the time course of the mandible movement is practically identical except for the local difference directly reflecting the consonantal (or rather syllable margin) gesture.

Given this agreement in the timing programs, the salient difference in the blade movement is rather remarkable. In particular, the stop gesture occurs earlier and starts moving

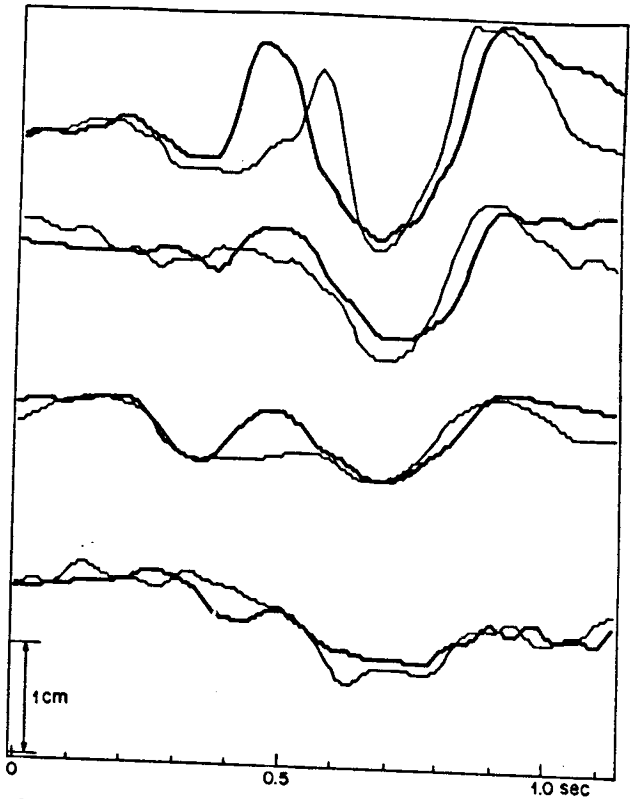

Fig. 3: Tongue blade movement (vertical position) comparing meaningless phrases 'bet aught' (thick lines) which was actually pronounced with a medial stop, and 'bed aught' (thin lines) pronounced with a medial tap by an American English speaker (female). The speaker was not given any instruction as to the manner of pronunciation, and the phrases were given in a list form. The curves represent, from top to bottom, tongue bade height, tongue dorsum height, mandible height, and tongue dorsum advancement, respectively. The two utterances are aligned in time to optimize the overall comparison for the curves except local deviations.

toward the vowel considerably earlier. The peak of the consonantal constriction (at the position of the pellet) occurs 100 msec or more earlier.

There are two categories of possible explanations. One is that the two gestures are, with respect to physiological realization, basically different. Different muscles are involved, perhaps associated with different neuromuscular and physical time constants, so that even the same time pattern of motor commands at the cortical level results in such a timing difference. Or perhaps the same muscle is used via different physiological cotrol mechanisms. Particularly, if the tapping gesture makes use of a more peripheral loop that elicits a response to some subtle change in a specific (perhaps unobserved) part of the articulator, as opposed to the stop gesture that is more or less under cortical control movement by movement, this rather qualitative difference might be plausible.

Another explanation may be that somehow the motor programming manifesting different syllable or foot structures has to be numerically different (perhaps only) in its timing configuration.

The former would suggest that the stop-flap (so-called) distinction is inherently discrete. The latter may imply that such allophonic variation is a continuous phenomenon; salient contrast evokes discretely or symbolically different perception or transcription, but depending on the context, particularly quantitatively specified phonetic parameters such as degree of emphasis or utterance speed, there may be intermediate cases from a phonetic point of view. The fact that some (even phonetically experienced) native speakers are not comfortable identifying the "stop-flap" distinction may suggest that the latter is the case. The recent study of formant characteristics of /l/-allophones by Sproat and Borowsky [1987] also seems to suggest the continuum of allophonic variation, refuting Halle and Mohanan's proposal [1985], and Borowsky's resyllabification theory basically seems promising in accounting for such phenomena.

4.4. An Elastic Model of Timing

As we have seen above, the temporal organization of speech for many purposes should be viewed as a multidimensional structure. If we obtain an effective descriptive system that uses approriate structural units for duration assignment, or events for defining timing and time intervals, then we may be able to represent in each of the dimensions the timing of each event by a linear model. That is, each time interval between contiguous events may be computed as a superposition of components due to different segmental and prosodic contributions as independent factors.

The idea of using a string of springs as a model of speech timing, or more specificallly of segmental durations, is not new. In particular, Jane Gaitenby [1965] at Haskins Laboratories discussed her data quite early using the concept of "elastic word" (see Lehiste [1980] for a review). What I would like to discuss here is a general model to describe the prosodic modulation of timing patterns of certain articulatory events. We can interpret time intervals among such events to derive durations of segmental units, whether phonemic, demisyllabic or syllabic.

After having determined the timing of each event in the time course of an utterance, we then will have to derive time functions of physical parameters, such as formant frequencies or tongue height, by looking up the inherent or segmental properties, static or dynamic as appropriate. As an example of such time function derivation processes, we may consider the case of pitch contours out of abstract tone specifications in Pierrehumbert's intonation work (see for its implementation as a synthesis rule system, Anderson *et al.* [1984]), or the prediction of velum movement patterns in Vaissiere's work discussed above.

Let us start with a simple example. Fig. 4 shows a single-dimension temporal model represented by a string of elastic springs. The length of each spring represents the time interval between two speech events to be observed, which are represented by joining points (circles) between congtiguous springs. The j-th spring is compressed or stretched deviating from its natural length x_{oj}, in response to the external force F. The extent of the response depends on the inherent stiffness k_j of the spring. Let us call the external force "prosodic force", because it is the cause of prosodic modulation in our model. The speed of utterance is directly related to the value of this prosodic force. Since the elastic system is superpositionally linear, all the increments of event intervals are proportional to the force. Also, I should emphasize here that the use of springs in the model does not imply uniform compression or stretching of the speech structure within the unit that is represented by a spring. Each spring represents only the interval between each adjacent pair of selected events.

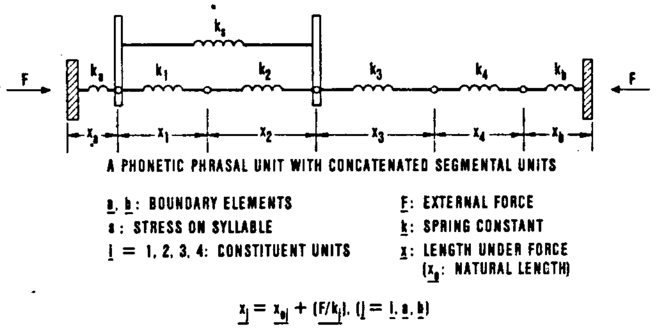

Fig. 4: A spring system composed for a simple phrasal unit (single dimension).

This model can be effective not only for representing the time interval distribution of an utterance as a whole, but also for the real time process of uttering a sentence, as far as the force is fixed throughout the utterance. This is so because in order to determine the conditions of the part of an utterance to the left (*i. e.* the past) of any joint point, we do not have to know the structure to its right (future). The boundary conditions for any substructure defined by the joint points at its ends are completely specified by the force applied to the end points. In this sense, considering the external force rather than the end positions as the boundary condition for the spring system is a crucial difference, even though mathematically the two ways of specifying boundary conditions are exactly equivalent.

We now need to devise a scheme to represent prosodic modulation. We would like to maintain that the quantities xo and k are inherent to the type of segment, or some projection of it onto a particular plane. We assume that the effect of stress is represented by a parallel spring, attached to the corresponding substring of segments representing a unit such as syllable, foot or word, to which such a stress specification is attached (see Fig. 4). This additional spring, which we may call a prosodic spring, has, as its inherent properties, natural length xo and stiffness k, representing the nature of the prosodic effect, such as the degree of stress. We may assume here that generally prosodic control is represented taking compression as the positive sense of the force. Thus a certain amount of external compression force is assumed for a neutral situation, and a relaxation or expansion of a segmental spring occurs when a parallel spring counterbalances part of the external force.

Another salient effect of prosodic modulation is the phrasal effect, in particular, phrase final lengthening. We represented in Fig. 4 such a boundary effect by adding virtual boundary springs, *a* and *b*, which are added to the segmental strings in series. These durational values are to be absorbed into the durational values of adjacent segmental units when we interpret the spring configuration as the temporal pattern of an utterance.

We can represent a hierarchical phrasal structure by embedding substructures in a larger spring system [Fujimura 1986d]. We need at least one level of phonetically motivated phrasal level to allow control of the prosodic status over the stretch of a syntactic phrase higher than words.

We studied this issue using icebergs (see *supra)* as the time marking events [Fujimura 1986a]. Fig. 5 illustrates a comparison of two utterances of the same word string, 'twenty two plus seven times four', distinguished by different phrasings corresponding to different arithmatic values. In this figure, each articulatory event (shown by a horizontal bracket) is plotted at the horizontal position representing its average timing, and at the vertical position representing the timing difference between the two utterances.

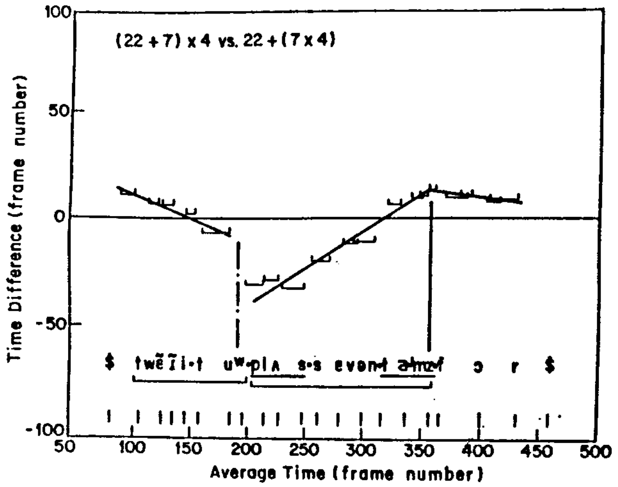

Fig. 5: A comparison of timings of corresponding events (iceberg-like movement patterns) in two utterances: along abscissa, average time for each event (shown by upward bracket), along ordinate, timing difference for each event between the two utterances. Note the time scales (in frame numbers) are different between the two axes. The frame interval is about 8 msec.

Tʰ. pattern of timing difference demonstrates a piecewise linear change. This means that the ratio of the interval difference to the average time interval, *i. e.* the percentage of interval variation between the pair, was constant over each phrase but varied from one phrase to another. The local utterance speed, or equivalently the prosodic force, was varied in a manner representable by a parallel phrasal spring. Each breakpoint of straight lines indicates where a phrase boundary occurred in the sentence in question. Actually, the piecewise linear change as seen in this figure indicates that there are additional constraints imposed on the constants of the constituent springs, within the general model discussed above. At any event, this observation seems to hold for all other similar utterance pairs. Furthermore, when we compare a sentence with a contrastive emphasis placed on different words, as discussed above, similar piecewise linear patterns obtain. In such a case, the emphasized word behaves as a prosodic phrase.

In my opinion, the traditional acoustic events for timing measurements are quite useful and reliable, but do not reveal some of the important characteristics of temporal organization. I suspect strongly that events do not occur synchronously among different articulatory dimensions, as I have discussed in previous papers, because movement patterns in each dimension can reflect articulator specific temporal constraints. Some aspects of the asynchronism are probably crucial for understanding the basic nature of the phonological/phonetic structure of speech [Fujimura 1981,1986; Allwood & Scully 1982; Scully and Allwood 1985]. We need to measure articulatory events, as well as voicing control, for different articulators simultaneously. The x-ray microbeam provides us a good means to obtain useful and rather comprehenssive data.

As I mentioned before, the complex spring model is useful for describing the utterance process as long as the following two conditions are met:
(1) The prosodic force is not altered in
the middle of an integral utterance unit,

(2) The change of plan takes place
by modifying the substructure
that connects to the past part of the utterance
only through a single node.

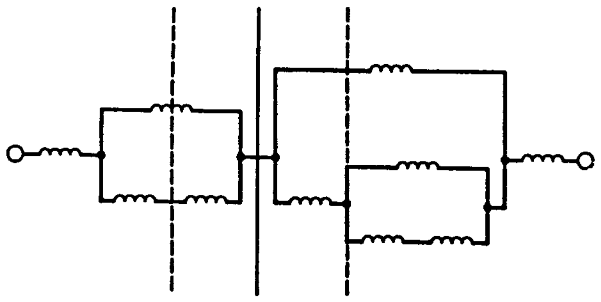

Fig. 6: A complex spring model (single dimension). Any of the solid vertical solidlines, but not broken lines, can mark a point in time as a boundary between past and future in motor programming.

Thus in Fig. 6, the solid vertical line can separate the past from the future, but broken vertical lines cannot. Time must jump from a single joint node to the next, as the motor program is sent for utterance execution. This suggests that the cognitive program controlling speech utterance is prepared as a sequence of phonetic phrasal units, formed into a simple concatenative linear string (*cf* Sternberg *et al.* [1978, 1980]). Within each phrasal unit, there can be any complex substructure involving parallel springs for prosodic modulations, but the specification of all such substructures within each phrasal unit must be complete before it gets started as an utterance.

Fig. 6 pertains only to one of the articulatory dimensions. The discussion above suggests also that the linkage among different dimensions must be solid at the phrase boundaries that serve for demarcating the motor program execution units.

5. Discourse and Intonation

In the tradition of linguistics, the sentence has played the most crucial role in the descriptive structure, defining the domain of syntax most successfully. Without delineating linguistic phenomena on the basis of the concept of sentence, the present achievement of descriptive theory could not have been imagined. This does not mean, however, that phenomena we encounter in speech research are all contained within the bounds of sentences. The more realistic we become in handling speech signals, the more severe we find the constraints. Recent research efforts have not overlooked these constraints. There are emerging findings which attempt at an ambitious challenge along this direction, even though, needless to say, it is very difficult to explore a rigorous theoretical approach, once we step out of the well-proved shelter of syntactic theory. These efforts are representatively characterized by a combination of training in artificial intelligence with the semantic, syntactic, phonological and phonetic as well as psychological disciplines. From the phonetic point of view, recent progress in intonation studies is particularly relevant, in many cases in connection with emerging thoughts about human linguistic performance and computational parsing of sentences.

Discussion of semantic references led AI researchers to the discovery of hierarchical block structures in the organization of discourse materials [Grosz 1977]. Recently, Hirschberg and Pierrehumbert [1986] (also Hirschberg [1987]) have shown that the voice pitch contours, when represented properly according to Pierrehumbert's descriptive framework, reveal the domains of such block structures. Silverman [1987] in his PhD dissertation corroborates this point, discussing related issues with extensive data from systematically controlled perceptual experiments. It is plausible that speech signals in actual conversation do carry more information than that represented in written text, in the form of pitch (and other signal aspects such as voice quality modulation as well as intensity), which significantly helps the listener in parsing the sentences correctly. Marcus and Hindle [1983] proposing their D-theory of syntactic descriiption argues that intonation breaks play a crucial role in sentence parsing, even though traditional orthographic systems ignore such information and necessitate for the readers of text a more complex parsing strategy.

6. Concluding Remarks

Speech is a physical and behavioral manifestation of linguistic structures. As such its characteristics can be evaluated only with reference to the linguistic structure that underlies it. While speech signals convey information other than linguistic codes, and the boundary between linguistic and extra- or para-linguistic issues may not be clearcut, there is no question that the primary goal of speech research is to understand the relation between the units and organizations of linguistic forms to properties of speech signals that are uttered and perceived under different circumstances. For this goal to be achieved, it is imperative that we have an effective (probably not the only correct) theory and a feasible representation framework based thereon. In my opinion, we have not established the linguistic theory to satisfy this condition, even though we have seen remarkable progress in recent years in this field, and our understanding now is far better and more useful than it was a decade ago.

Furthermore, such theoretical endeavors must depend crucially on experimental and computational approaches, and is sensitive to the needs of industrial applications, just as, I might say, theoretical work in solid state physics is. Thus our work in speech synthesis from text, for example, can be affected immediately by any innovative development of the level theory [Kiparsky 1982] in morphology, nonlinear phonology, lexical semantics, syntactic subcategorization of verbs and so on, as well as discussions of temporal organization of articulatory events. On the other hand, there is no question that the theoretical discussion of the emerging tier/plane theory of phonological description must crucially depend on a rather accurate description of pitch contours, anatomy and physiology of articuratory systems, movement patterns of the tongue, the lips, the jaw, etc. along with a good linguistic insight and factual knowledge of a variety of natural languages, synchronic and diachronic. In addition, empirical results we obtain in engineering implementations of the theories do provide us with invaluable suggestions as to the future direction, as well as good motivations.

The more we learn about speech and language, the more strongly are we impressed by the depth of the human cognitive faculty.

7. Acknowledgment

I wish to acknowledge the benefit of daily discussions of relevant issues with all my colleagues in the Departments of Linguistics and of Artificial Intelligence Research, AT&T Bell Laboratories. In addition, Kim Silverman and David Talkin read through earlier versions of my manuscript and contributed many helpful comments. I am also grateful for many expositional corrections contributed by several readers, particularly Nina MacDonald and Joan Bachenko.

REFERENCES

Abbs, J.H.(1979), Speech Motor Equivalence: The Need for a Multi-Level Control, *Proc. 9th Int. Cong. Phonetic Sc., Vol II*, Copenhagen, Aug 6-11, 318-324.

Abbs, J.H. and Gracco, V.L.(1982); Motor Control of Multi-Movement Behaviors: Orofacial Muscle Responses to Load Perturbations of the Lips during Speech, *Soc. Neuroscience 8*, 282.

Abbs, J.H. and Gracco, V.L.(1983), Sensorimotor Actions in the Control of Multimovement Speech Gestures, in *Trends in Neuroscience 6*, 393-395.

Abbs, J.H. and Gracco, V.L.(in press), Control of Multimovement Coordination: Sensorimotor Mechanisms in Speech Motor Programming, *J. Motor Behavior*.

Allwood, E. and Scully, C.(1982), A Composite Model of Speech Production, *Proc. ICASSP '82, Vol. 2*, Piscataway N.J., IEEE Service Center, 932-935.

Anderson, M., Pierrehumbert, J.B. and Liberman, M.Y.(1984), Synthesis by Rule of English Intonation Patterns, *Proc. ICASSP '84, Vol. 1*, Piscataway N.J., IEEE Service Center,2.8.1-2.8.4.

Baer, T., Sasaki, C. and Harris, K.(eds.),(1987), *Laryngeal Function in Phonation and Respiration*, San Diego, College Hill Press.

Beckman, M.E. and Pierrehumbert, J.B.(1986), Intonational Structure in Japanese and English, in *Phonology Yearbook 3*, C. Ewen and E. Anderson(eds.), Cambridge, Cambridge U. Press,255-309.

Bell-Berti, F. and Harris, K.S.(1982), Temporal Patterns of Coarticulation: Lip Rounding, *J. Acoust. Soc. Am. 71*, 449-454.

Bless, D.M. and Abbs, J.(eds.)(1983), *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, San Diego, College Hill Press.

Blumstein, S. E. and Stevens, K. N.(1979), Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characterisitcs of Stop Consonants, *J. Acoust. Soc. Am. 66*, 1001-1017.

Borowsky, T.J.(1986), *Topics in the Lexical Phonology of English*, PhD Diss., Dept Linguistics, U. Massachussets, Amherst.

Browman, C.P.(1980), Rules for Demisyllable Synthesis Using LINGUA, a Language Interpreter, *Proc. ICASSP '80, Vol 2*, Piscataway N.J., IEEE Service Center,561-164.

Browman, C.P. and Goldstein, L.M.(1985), Dynamic Modeling of Phonetic Structure, in *Phonetic Linguistics -- Essays in Honor of Peter Ladefoged*, V. A. Fromkin(ed.), New York, Academic Press,35-53.

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

Childers, D.G., Smith, A.M. and Moore, G.P.(1984), Relationships between Electroglottograph, Speech and Vocal Cord Contact, *Folia Phoniatrica 36*, 105-118.

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

Coker, C.H.(1968), Speech Synthesis with a Parametric Articulatory Model in *Speech Symposium, Kyoto 1968*, reprinted in *Speech Synthesis*, J. L. Flanagan and L. R. Rabiner(eds.), Stroudsburg, Penn, Dowden-Hutchinson, Ross,135-139.

Conrad, W.(1987), Simplified One Mass Model with Supraglottal Resistance: A Testable Hypothesis, in *Laryngeal Function in Phonation and Respiration*, T. Baer, C. Sasaki and K. Harris(eds.), San Diego, College Hill Press, 320-338.

Cooper, F.S., Delattre, P., Liberman, A.M., Borst, J. and Gerstman, L. (1952), Some Experiments on the Perception of Speech Sounds, *J. Acoust. Soc. Am. 24*, 579-606.

Davis, P.J. and Nail, B.S. (to appear), The Sensitivity of Laryngeal Epithelial Receptors to Static and Dynamic Forms of Mechanical Stimulation, in *Voice Production*, O.Fujimura(ed.), New York, Rave Press.

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

DeLuca, C.J.(1975), A Model for a Motor Unit Train Recorded during Constant Force Isometric Contractions, *Biol. Cybern. 19*, 159-167.

Dinnsen, D. A. and Charles-Luce, J.(1984), Phonological Neutralization, Phonetic Implementation and Individual Differences, *J. Phonetics 12*, 49-60.

Dixon, N.R. and Maxey, H.D.(1968), Terminal Analog Synthesis of Continuous Speech Using the Diphone Method of Segment Assembly, *IEEE Trans. Audio electroacoustics 16*, 40-50.

Edwards, J.(1985), *Mandibular Rotation and Translation during Speech*, PhD Diss., City U. of New York, Dept. Speech Hear. Scs.

Eek, A.(1973), Observations in Estonian Plalatalization: An Articulatory Study, *Estonian Papers in Phonetics*, 18-36.

Eek, A.(ed.)(1978), Studies on Accent, Quantity, Stress, Tone: Papers of the Symposium, Tallinn, Nov. 1978, *Estonian Papers in Phonetics*.

Eek, A. and Remmel, M.(1974), Context, Contacts and Duration: Two Results concerning Temporal Organization, *Preprints of the Speech Communication Seminar, Stockholm, Aug. 1-3, 1974*, 187-192.

Fant, G.(1956), On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies, in *For Roman Jakobson*, M. Halle, H. G. Lunt, H. McLean and C. C. van Schooneveld(eds.), The Hague, Mouton,109-120.

Fant, G.(1960), *Acoustic Theory of Speech Production*, The Hague, Mouton.

Fant, G.(1983), Preliminaries to Analysis of the Human Voice Source, *KTH-QPSR [illegible] 1982), Royal Inst. Technol.*, Sweden,1-27.

Fant, G.(1983a), T[illegible] [illegible]oice Source: Acoustic Modeling, *KTH-QPSR 4 (1982), [illegible] Inst. Technol.* Sweden,28-48.

Fant, G., Ondračkova, J., Lindqvist, J. and Sonesson, B.(1966), Electrical Glottography, *KTH-QPSR 4 (1982), Royal Inst. Technol.* Sweden,15-21.

Firth, I. M.(1986), Modal Analysis of the Vocal Tract, *J. Acoust. Soc. Am. 80, Suppl 1*, S97.

Fisher, W., Doddington, G. and Goudie-Marshall, K.(1986), The DARPA Speech Recognition Database: Specifications and Status, *Proc. the DARPA Speech Recognition Workshops*, Feb. (1986), Washington, D. C., Science Applications International Corp.

Flanagan, J.L. and Landgraf, L.(1968), Self-Oscillating Source for Vocal- Tract Synthesizers, *IEEE Trans. Audio-Electronics*, 57-64.

Flanagan, J.L. and Ishizaka, K.(1976), Automatic Generation of Voiceless Excitation in a Vocal Cord/Vocal-Tract Speech Synthesizer, *IEEE Acoust. Signal Speech Processing 24*, 163-170.

Fletcher, S.G., McCutcheon, M.J. and Wolf, M.S.(1975), Dynamic Palatometry, *J Speech Hear. Res. 18*, 812-819.

Fourakis, M. (1984), Should Neutralization be Redefined?, *J. Phonetics 12*, 291-296.

Fourakis, M. and Iverson, G.K.(1984), On the 'Incomplete Neutralization' of German Final Obstruents, *Phonetica 41*, 140-149.

Fourcin, J. and Abberton, E.(1977), Laryngographic Studies of Vocal Cord Vibration, *Phonetica 34*, 313-315.

Frances, W.N. and Kučera, H.(1982), *Frequency Analysis of English Usage*, Boston, Houghton Mifflin Co.

Fromkin, V.A.(ed.)(1985), *Phonetic Linguistics, Essays in Honor of Peter Ladefoged*, Orlando, Florida, Academic Press.

Fujii, I., Fujimura, O. and Kagaya, R.(1971), Dynamic Palatography by use of a Computer and an Oscilloscope, *Proc. 7th Int. Cong. Acoust., Vol 3*, 113-116.

Fujimura, O.(1961), Bilabial Stop and Nasal Consonants: A Motion Picture Study and its Acoustical Implications, *J Speech Hear. Res. 4*, 233-2247.

Fujimura, O.(1970), Current Issues in Phonetics, *Studies in General and Oriental Linguistics*, R. Jakobson and S. Kawamoto(eds.), Tokyo, TEC Co., 109-130.

Fujimura, O.(1976), Syllable as Concatenated Demisyllables and Affixes, *Proc. Acoust. Soc. Am. 59, Suppl 1*, S55.

Fujimura, O.(1977), Recent Findings on Articulatory Processes -- Velum and Tongue Movements as Syllable Features, in *Articulatory Modeling and Phonetics*, R.Carre, R.Descout and M.Wajskop(eds.), Grenoble, GALF Groupe de la Communication Parlee,115-126.

Fujimura, O.(1979), Physiological Functions of the Larynx in Phonetic Control, *Current Issues in the Phonetic Sciences*, H. and P. Hollien(eds.), Amsterdam, John Benjamins,129-164.

Fujimura, O.(1980), Modern Methods of Investigation in Speech Production, *Phonetica 37*, 38-54.

Fujimura, O.(1981), Temporal Organization of Articulatory Movements as a Multidimensional Phrasal Structure, *Phonetica 38*, 66-83, corrected version in *Proc. Symp. Acoustic Phonetics and Speech Modeling, Part 2, Paper 5*, Inst. Defense Analysis, Princeton N.J..

Fujimura, O.(1981a), Body-Cover Theory of the Vocal Fold, in *Vocal Fold Physiology*, K. N. Stevens and M. Hirano(eds.), Tokyo, U. Tokyo Press,271-281.

Fujimura, O.(1984), The Role of Linguistics for Future Speech Technology, *Ling. Soc. Am. 104*, June 1984.

Fujimura, O.(1986), Relative Invariance of Articulatory Movements: An Iceberg Model, in *Invariance and Variability in Speech Processes*, J. S. Perkell and D. H. Klatt(eds.), Hillsdale N.J., Lawrence Erlbaum,226- 242.

Fujimura, O.(1986a), Temporal Organization of Articulatory Movements - A Multidimensional Complex Spring Model for Running Speech, *Proc. Acoust. Soc. Am. 80, Suppl 1*, S97.

Fujimura, O.(1986b), Evaluating the Task Dynamics Model, *J. Phonetics 14*, 105-108.

Fujimura, O.(1986c), A Model of Temporal Organization of Articulatory Gestures -- An X-ray Microbeam Study *Folia Phoniatrica 38*, 298, (Abst.).

Fujimura, O.(1986d), A Linear Model of Speech Timing, in *Ilse Lehiste Puhendusteos*, R.Channon and L.Shockey(eds.), Dordrecht, Holland, Foris.

Fujimura, O.(ed.)(to appear), *Voice Production - Proc. 5th Vocal Fold Physiology Conference Jan. 1987(tentative title)*, New York, Raven Press.

Fujimura, O. and Lindqvist, J.(1971), Sweep-Tone Measurements of Vocal Tract Characteristics, *J. Acoust. Soc. Am. 49*, 541-558.

Fujimura, O., Ishida, Y. and Kiritani, S.(1973), Computer-Controlled Radiography for Observation of Movements of Articulatory and other Human Organs, *Comp. Biology & Med. 3*, 371-384.

Fujimura, O., Tatsumi, I.F. and Kagaya, R.(1973a), Computational Processing of Palatographic Patterns, *J. Phonetics 1*, 47-54.

Fujimura, O., Miller, J.E. and Kiritani, S.(1977), A Computer-Controlled X-Ray Microbeam Study of Articulatory Characteristics of Nasal Consonants in English and Japanese, *Proc. 9th Int Congr. on Acoustics, Contributed Papers 1*, Madrid, 461.

Fujimura, O., Macchi, M.J. and Lovins, J.B.(1977a) Demisyllables and Affixes for Speech Synthesis, *Proc. 9th Int Congr. on Acoustics, Contributed Papers 1*, Madrid, 513.

Fujimura, O. and Lovins, J.(1978), Syllables as Concatenative Phonetic Units, in *Syllables and Segments*, A. Bell and J. B. Hopper(eds.), Amsterdam, North Holland, 107-120. An unabridged version available from Indiana U. Ling. Club.

Fujimura, O. and Kakita, Y.(1979), Remarks on Quantitative Description of the Lingual Articulation, in *Frontiers of Speech Communication Research*, S. Öhman and B. Lindblom(eds.), London, Academic Press,17-24.

Fujisaki, H.(1977), Functional Models of Articulatory and Phonatory Dynamics, in *Proc. US-Japan Sem. Dynamic Aspects of Speech Production*, M. Sawashima and F. S. Cooper(eds.), Tokyo, U. Tokyo Press,347-366.

Fujisaki, H.(1983), Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing, in *The Production of Speech*, P. F. MacNeilage(ed.), New York, Springer-Verlag, 39-56.

Gårding, E. (1983), A Generative Model of Intonation, in *Prosody: Models and Measurements*, A. Cutler and D. R. Ladd(eds.), New York, Springer-Verlag, 11-26.

Gaitenby, J.H.(1965), The Elastic Word, *Status Rep. Speech Res. 2*, New Haven, Haskins Laboratories, 3.1-3.11.

Grosz, B.J.(1977), *The Representation and Use of Focus in Dialogue Understanding*. Technical Rept 151, Menlo Park CA, SRI International.

Haken, H.(1977), *Synergetics* (Third Edition 1983), Heidelberg, Springer Verlag.

Halle, M.(1983), On Distinctive Features and their Articulatory Implementation, *Nat. Lang. Linguistic Theory 1*, 91-105.

Halle, M.(1985), Speculations about the Representation of Words in Memory, in *Phonetic Linguistics*, V. A. Fromkin(ed.), New York, Academic Press, 101-114.

Halle, M. and Mohanan, K.(1985), Segmental Phonology of Modern English, *Linguistic Inquiry 16*, 57-116.

Halle, M. and Vergnaud, J.-R.(1987), Stress and the Cycle, *Linguistic Inquiry 18*, 45-84.

Haraguchi, S.(1975), *The Tone Pattern of Japanese: An Autosegmental Theory of Tonology*, PhD Diss., Dept. Linguistics & Philosophy, MIT, Cambridge MA.

Hardcastle, W.(1972), The Use of Electropalatography in Phonetic Research, *Phonetica 25*, 192-215.

Hardcastle, W.J.(1974), Instrumental Investigation of Lingual Activity during Speech: A Survey, *Phonetica 29*, 129-157.

Hardcastle, W.(1984), New Methods of Profiling Lingual Palatal Contact Patterns with Electropalatography, *WP Phonetics Lab., U. Readings, No 4*, 1-40.

Hattori, S.(1961), Prosodeme, Syllable Structure and Laryngeal Phonemes, in *Studies in Descriptive and Applied Linguistics, Bulletin of the Summer Inst.Linguistics, 1*, Internat. Christian Univ., Tokyo, 1-27.

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

Henke, W.L.(1966), *Dynamic Articulatory Model of Speech Production Using Computer Simulation*, PhD. Diss., Dept. Linguistics & Philosophy, MIT, Cambridge, Mass.

Hirano, M.(1977), Structure and Vibratory Behavior of the Vocal Folds, in *Dynamic Aspects of Speech Production*, M. Sawashima and F. S. Cooper(eds.), Tokyo, U. Tokyo Press.

Hirano, M. and Ohala, J.(1969), Use of Hooked-Wire Electrodes for Electromyography of the Intrinsic Laryngeal Muscles, *J. Speech Hear. Res. 12*, 362-373.

Hirose, H. and Kiritani, S.(1985), A Kinesiological Study of Labial Articulatory Movements in Ataxic Patients, *Ann. Bull. U. Tokyo RILP 19*, 201-208.

Hirschberg, J. (1987), Uses of Intonational Cues in Discourse Studies, *Proc. TINLAP-87*. 86-91.

Hirschberg, J. and Pierrehumbert, J.B.(1986), The Intonational Structuring of Discourse, *Proc. 24th Meeting Ass. Comp. Linguistics*, New York, Columbia Univ.,136-144.

Holmes, J.N. and Mattingly, I.G.(1964), Speech Synthesis by Rule, *Language & Speech 7*, 127-143.

Honda, K.(1983), Variability Analysis of Laryngeal Muscle Activity, in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, I.Titze and R.Scherer(eds.), Iowa, Iowa U. Press,127-137.

Honda, K., Kiritani, S., Imagawa, J. and Hirose, H.(1985), High-Speed Digital Recording of Vocal Fold Vibration Using a Solid-State Sensor, *Ann. Bull. U. Tokyo RILP 19*, 47-53.

Horiguchi, S. and Bell-Berti, F.(1984), The Velotrace: A Mechanical Device for Tracking Velar Position, *Paper Presented at the Meeting of American Cleft Palate Assoc.*, Seatle, WA.

Houde, R.(1967), *A Study of Tongue Body Motion During Selected Speech Sounds*, PhD Diss., U. Michigan, *Speech Comm. Res. Lab. Monograph 2 (1968)*, Los Angeles.

House, A.S. and Stevens, K.N.(1956), Analog Studies of the Nasalization of Vowels, *J. Speech Hear. Dis. 21*, 218-232.

Hughes, O.M. and Abbs, J.H.(1976), Labio-Mandibular Coordination in the Production of Speech: Implications for the Operator of Motor Equivalence, *Phonetica 33*,, 199-221.

Inkelas, S., Leben, W. and Cobler, M.(1987), The Phonology of Intonation in Hausa, *NELS 17*, North-Eastern Linguistic Society.

Ishizaka, K. and Matsudaira, M.(1968), What Makes the Vocal Cords Vibrate?, *Proc. 6th Internat. Cong. Acoustics, Vol II*, Y. Kohasi(ed.), New York, Elsevier,9-12.

Jakobson, R., Fant, C.G. and Halle, M.(1951), *Preliminaries to Speech Analysis*, Cambridge, MIT Press (Third Edition 1963).

Kahn, D. (1976) *Syllable-Based Generalizations in English Phonology*, PhD Diss., Dept. Linguistics & Philosophy, MIT, Cambridge, Mass. New York, Garland Publishing, Inc. (1980).

Kakita, Y., Hirano, M. and Ohmaru, K.(1981), Physical Properties of the Vocal Fold Tissue: Measurements on Excised larynges, *Vocal Fold Physiology*, K. N. Stevens and M. Hirano(eds.), Tokyo, U. Tokyo Press, 377-397.

Kakita, Y., Fujimura, O. and Honda, K.(1985), Computation of Mapping from Muscular Contraction Patterns to Formant Patterns in Vowel Space, in *Phonetic Linguistics*, V. A. Fromkin(ed.), New York, Academic Press,133-144.

Kaneko, T., Uchida, K.,Suzuki, H., Komatsu, K., Kanesaka, T., Kobayashi, N. and Naito, J.(1981), Ultrasonic Observations of Vocal Fold Vibration, in *Vocal Fold Physiology*, K. N. Stevens and M. Hirano(eds.), Tokyo, Univ.Tokyo Press,107-117.

Kean, M. (1975), *The Theory of Markedness in Generative Grammar*, PhD Diss., Dept. Linguistics & Philosophy, MIT, Cambridge MA.

Keating, P. A.(1985), Universal Phonetics and the Organization of Grammars, in *Phonetic Linguistics*, V. A. Fromkin(ed.), New York, Academic Press, 115-132.

Kelso J.A.S., Saltzman E.L. and Tuller, B.(1986), The Dynamical Perspective on Speech Production, *J. Phonetics 14*, 29-59.

Kelso, J.A.S., Saltzman, E.L. and Tuller, B.(1986a), Intentional Contents, Communicative Context, and Task Dynamics, a Reply to Commentators, *J. Phonetics 14*, 171-196.

Kent, R.D. (1983), The Segmental Organization of Speech, in *The Production of Speech*, P. F. MacNeilage(ed.), New York, Springer-Verlag, 57-90.

Kiparsky, P.(1982), From Cyclic Phonology to Lexical Phonology, *The Structure of Phonological Representation, Part I*, H. van der Hulst and N. Smith(eds.), Dordrecht, Holland, Foris Publication,131-176.

Kirikae, I.(1943), A Study on the Vibration of the Human Vocal Cords in Phonation and the Timing Relations of the Glottal Opening-Closure by the Use of a Laryngeal Stroboscopic Motion Picture Technique, *J. Japan Otorhinolaryngology 49*, 236-268.

Kiritani, S., Itoh, K. and Fujimura, O.(1975), Tongue-Pellet Tracking by a Computer Controlled X-Ray Microbeam System, *J. Acoust. Soc. Am. 57*, 1516-1520.

Kiritani, S., Miyawaki, K. Fujimura, O. and Miller, J. E.(1976), A Computational Model of the Tongue, *Ann. Bull. U. Tokyo RILP 10*, 243-251.

Kiritani, S., Tateno, Y. and Iinuma, T.(1977), Computer Tomography of the Vocal Tract, in *Dynamic Aspects of Speech Production*, M. Sawashima and F. S. Cooper(eds.), Tokyo, U. Tokyo Press, 203-208.

Kiritani, S. Imagawa, H. and Hirose, H.(to Appear), High-Speech Digital Image Recording for the Observation of Vocal Cord Vibration, in O.Fujimura(ed.), *Voice Production*, Raven Press.

Klatt, D.H.(1978), Synthesis by Rule of Consonant-Vowel Syllables, *Proc. Acoust. Soc. Am. 64, Suppl 1*, S114.

Klatt, D.H.(1979), Synthesis by Rule of Segmental Durations in English Sentences, in *Frontiers of Speech Communication Research*, B. Lindblom and S. Ohman(eds.), New York, Academic Press, 287-300.

Kohler, K J. and van Dommelen, W. A. (1986), Prosodic Effects on Lenis/Fortis Perception: Preplosive F0 and LPC Synthesis, *Phonetica 43*, 70-75.

Kozhevnikov, V.A. and Chistovic, L.A.(1965), *Rech: Artikulyatsia I Vospriyatiye*, Nauka, Moscow, Leningrad.

Kupin, J. J.(1979), *Tongue Twisters as a Source of Information about Speech Production*, PhD Diss., Storrs CT, U. Conn.

Ladd, R.(1986), Intonational Phrasing: the Case for Recursive Prosodic Structure, *Phonology Yearbook 3*, C. Ewen and E. Anderson (eds.), Cambridge; Cambridge U. Press, 311-340.

Ladefoged, P., Cochran, A. and Disner, S.(1977), Laterals and Trills, *J. Int. Phonetic Assoc. 7*, 46-54.

Laferriere, M. and Zue, V.W,(1977), Flapping Rule in American English: An Acoustical Study, *J. Acoust. Soc. Am. 61, Suppl 1* S31.

Lassen, N.A. and Larsen, B.(1980), Cortical Activity in Left and Right Hemisphere during Language Related Brain Functions, *Proc. 9th Int. Cong. Phonetics Sci., Vol III*, Copenhagen, Aug 6-11, 137-150.

Lehiste, I. (1970), *Suprasegmentals,* Cambridge, Mass, MIT Press.

Lehiste, I.(1980), Phonetic Manifestation of Syntactic Structure in English, *Ann. Bull. U. Tokyo RILP 14* 1-28.

Liberman, A.M., Delattre, P.C., Cooper, F.S. and Gerstman, L. J.(1954), The Role of Consonant-Vowel Transitions in the Perception of Stop and Nasal Consonants, *Psychology Monographs 68*, 1-13.

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P. and Cooper, F.S.(1959), Minimal Rules for Synthesing Speech, *J. Acoust. Soc. Am. 31*, 1490-1499.

Liberman, M.Y.(1975) *The Intonational System of English*, PhD Diss., Dept. Linguistics & Philosophy, MIT, Cambridge, Mass.

Liberman, M.Y. and Prince, A.(1977), On Stress and Linguistic Rhythm, *Linguistic Inquiry 8*, 249-336.

Lindblom, B.(1964), Dynamic Aspects of Vowel Articulation, *Proc. 5th Int. Cong. Phonetic Science*, Munster, 387-388.

Lindblom, B.(1968), *On the Production and Recognition of Vowels*, PhD Diss., Lund Univ.

Lindblom, B.(1983), Economy of Speech Gestures, in *The Production of Speech*, P. F. MacNeilage(ed.), New York, Springer-Verlag,217-246.

Lindblom, B., Lubker, J. and Gay, T.(1979), Formant Frequencies of some Fixed-Mandible Vowels and a Model of Speech Motor Programming by Predictive Simulation, *J. Phonetics 7*, 147-161.

Macchi, M.J.(1980), A Phonetic Dictionary for Demisyllable Speech Synthesis, *Proc. ICASSP '80, Vol 2*, Piscataway NJ, IEEE Service Center, 565-567.

Macchi, M.J.(1985), *Segmental and Suprasegmental Features and Lip and Jaw Articulators*, PhD Diss., Dept. Linguistics, New York Univ.

MacNeilage, P.F.(ed.) (1983), *The Production of Speech*, New York, Springer-Verlag.

MacNeilage, P.F.(1985), Serial-Ordering Errors in Speech and Typing,in *Phonetic Linguistics*, V.A.Fromkin(ed.), Orlando, Florida, Academic Press.

Maeda, S.(1976), *A Characterization of American English Intonation*, PhD Diss., MIT, Cambridge MA.

Maeda, S.(1983), *Correlats Acoustiques de la Nasalization des Voyelles: Une Étude de Simulation*, Centre National D'études des Télecommunications, Centre des Recherches de Lannion, Rept. VII.

Marcus, M. and Hindle, D.(1983), D-Theory: Talking about Talking about Trees, *Proc. 21st Annual Meeting of the Ass. Comp. Linguistics*.

McCawley, J.D.(1968), *The Phonological Component of a Grammar of Japanese*, The Hague, Mouton.

Mermelstein, P.(1973), Articulatory Model for the Study of Speech Production, *J. Acoust. Soc. Am. 53*, 1070-1082.

Miyawaki, K.(1972), *A Study of Lingual Articulation by Use of Dynamic Palatography*, M.A. Thesis, Dept. Linguistics, U. Tokyo.

Müller, E., Abbs, J, Kennedy, J. and Larson, C. (1977), Significance of Perioral Biomechanics to Lip Movements during Speech, *Am. Speech Lang. Hear. Assoc.*

Nadler, R.D., Abbs, J.H., Fujimura, O.(1987), Speech Movement Research Using the New X-Ray Microbeam System, *Proc. 11th International Congress of Phonetic Sciences, Tallinn*, Paper Se 11.4.

Nelson, W.L.(1983), Physical Principles for Economies of Skilled Movements, *Biol. Cybernetics 46*, 135-147.

Nishinuma, Y. and Rossi, M.(1981), Automatisation of Prosodic Analysis in French, *Study of Sounds (The Phonetic Society of Japan) XIX*, 155-169.

Nooteboom, S. G. and Terken, J. M. B.(1982), What Makes Speakers Omit Pitch Accents? An Experiment, *Phonetica 39*, 317-336.

Ohala, M.(1983), "The Machine as an Addressee: When Paralinguistics Fails, *Abstracts of the Tenth International Congress of Phonetic Sciences*, Dordrecht, Foris Publication,428.

Öhman, S.E.G. (1967), Numerical Model of Coarticulation, *J. Acoust. Soc. Am. 41*, 310-320.

Oka, D.K.(1980), *The Design and Test of a Ranging Transducer to Monitor Articulatory Movement during Speech Production*, M.S. Thesis, MIT.

Olive, J.P.(1980), A Scheme of Concatenating Units for Speech Synthesis, *Proc. ICASSP '80, Vol 2*, Piscataway N.J., IEEE Service Center,568-571.

Ostry, D. J., Keller, E. and Parush, A.(1983), Similarities in the Control of Speech Articulators and Limbs: Kinematics of Tongue Dorsum Movement in Speech, *J. Exp. Psychology: Human Percept. & Perf. 9*, 622-636.

Perkell, J.S.(1969), *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Analysis*, Res. Monograph No 53, Cambridge, Mass, MIT Press.

Perkell, J.S. and Nelson, W.L.(1982), Articulatory Targets and Speech Motor Control: A Study of Vowel Production, in *Speech Motor Control*, S. Grillner, B. Lindblom, J. Lubker, A. Persson(eds.), New York, Pergamon, 187-204.

Perkell, J. and Klatt, D.(eds.)(1986), *Invariance and Variability in Speech Processes*, Hillsdale NJ, Lawrence Erlbaum.

Pierrehumbert, J.(1980), *The Phonology and Phonetics of English Intonation*, PhD Diss., Dept. Linguistics & Philosophy, MIT, Cambridge MA.

Pierrehumbert, J. and Beckman, M.(in press), Japanese Tone Structure, *Linguistic Inquiry*.

Port, R.F. and O'Dell, M.L.(1985), Neutralization of Syllable-Final Voicing in German, *J. Phonetics 13*, 455-471.

Poser, J.P.(1984), *The Phonetics and Phonology of Tone and Intonation in Japanese*, PhD Diss., Dept. Linguistics & Philosophy, MIT, Cambridge MA.

Rome, J.A.(1964), An Artificial Pellet for Continuous Analysis of Speech, *MIT Res. Lab. Eletronics, Quart. Prog. Rep. 74*, 190-191.

Rosenberg, E.A., Rabiner, R.L., Wilpon, G.J. and Kahn, D.(1983), Demisyllable-Based Isolated Word Recognition System, *IEEE Trans. Acoust. Speech & Signal Process 31*, 713.

Rossi, M. and Autesserre, D. (1981), Movements of the Hyoid and the Larynx and the Intrinsic Frequency of Vowels, *J. Phonetics 9*, 233-249.

Rothenberg, M.(1981), Acoustic Interaction between the Glottal Source and the Vocal Tract, in *Vocal Fold Physiology*, K. N. Stevens and M. Hirano(eds.), Tokyo, U. Tokyo Press, 305-328.

Saito, S., Fukuda, H., Isogai, Y. and Ono, H.(1981), X-ray Stroboscopy, in *Vocal Fold Physiology*, K. N. Stevens and M. Hirano(eds.), Tokyo, U. Tokyo Press, 95-106.

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

Sawashima, M.(1976), Fiberoptic Observation of the Larynx and other Speech Organs, *Proc. US-JAP Sem. Dynamic Asp. Speech Production*, Tokyo, U. Tokyo Press, 31-46.

Sawashima, M. and Kiritani, S.(1985), Electro-Palatographic Patterns of Japanese /d/ and /r/ in Intervocalic Position, *Ann. Bull. U. Tokyo RILP 19*, 1-6.

Schönle, P.W., Wenig, P., Schrader, J., Grabe, K., Brockmann, E. and Conrad, B. (1983), Ein Elektromagnetisches Verfahren zur Simultanen Registrierung von Bewegungen im Bereich des Lippen-, Unterkiefer- und Zungensystems, *Biomed. Technik 28*, 263-267.

Scully, C. and Allwood, E. (1985), Production and Perception of an Articulatory Continuum for Fricatives in English, *Speech Communications 4*, 237-246.

Silverman, K.(1987), *The Structure and Processing of Fundamental Frequency Contours*, PhD Diss., Cambridge Univ.

Sivertsen, E.(1961), Segment Inventories for Speech Synthesis, *Language & Speech 4*, 27-89.

Shadle, C.(1985), *The Acoustics of Fricative Consonants*, PhD Diss., MIT.

Shibata, S., Ino, A. and Yamashita, S. (1979), *Teaching Articulation by Use of Electro-Palatography* (English translation 1982 available), Kokubunji, Tokyo, Rion Co., Ltd.

Smith, S.(1981), Research on the Principle of Electroglottography, *Folia Phoniatrica 33*, 105-114.

Sonies, B.C., Shawker, T.H., Hall, T.E., Gerber, L.H. and Leighton, S.B.(1981), Ultrasonic Visualization of Tongue Motion during Speech, *J. Acoust. Soc. Am. 70*, 683-686

Sonoda, Y. and Kiritani, S.(1976), Analysis of Tongue Point Movements by a Linear Second-Order System Model, *U. Tokyo, Ann. Bul. RILP 10*, 29-36.

Sproat, R. and Borowsky, T.(1987), On the Resyllabification of /l/ in English, *J. Acoust. Soc. Am. 81, Suppl. 1*, S67.

Stevens, K.N.(1960), Towards a Model for Speech Recognition, *J. Acoust. Soc. Am. 32*, 47-55.

Stevens, K.N.(1972), The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data, in *Human Communication: A Unified View*, E. E. David and P. B. Denes(eds.), New york, McGraw-Hill, 51-66.

Stevens, K.N.(1975), Physics of Laryngeal Behavior and Larynx Modes, *Phonetica 34*, 264-279.

Stevens, K.N.(1983), Design Features of Speech Sound Systems, in *The Production of Speech*, P. F. MacNeilage(ed.), New York, Springer Verlag,247-262.

Stevens, K.N.(to appear), Modes of Vocal Fold Vibration Based on a Two-Section Model, in *Voice Production*, O.Fujimura(ed.) New York, Raven Press.

Stevens, K.N. and Hirano, M.(eds.)(1981), *Vocal Fold Physiology* Tokyo, U. Tokyo Press.

Sternberg, S., Monsell, S., Knoll, R. and Wright, C.(1978), The Latency and Duration of Rapid Movement Sequences: Comparison of Speech and Typewriting, in *Information Processing in Motor Control and Learning*, G. Stelmach(ed.), New york, Academic Press.

Sternberg, S., Wright, C.E., Knowll, R.L. and Monsell, S.(1980), Motor Programs in Rapid Speech, Additional Evidence, in *The Perception and Production of Fluent Speech*, R. A. Cole(ed.), Hillsdale NJ, Lawrence Erlbaum, 507-534.

Teager, H.M.(1983), The Effects of Separated Air Flow on Vocalization, in *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, D. M. Bless and J. H. Abbs(eds.), San Diego, College-Hill Press, 124-141.

Thorsen, N.(1984), Variability and Invariance in Danish Stress Group Patterns, *Phonetica 41*, 88-102.

Thomas, T.J.(1985), *An Articulatory Model of Speech Production Including Turbulence*, PhD Diss., Cambridge Univ.

Titze, I.R.(1985), Mechanisms of Sustained Oscillation of the Vocal Folds, in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, I. R. Titze and R. C. Scherer(eds.), Denver, The Denver Center for the Performing Arts, 349-357.

Titze, I.R. and Talkin, D.T.(1979), A Theoretical Study of the Effects of Various Laryngeal Configurations on the Acoustics of Phonation, *J. Acoust. Soc. Am. 66*, 60-74.

Titze, I. and Scherer, R.(eds.)(1983), *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, Denver, Denver Center of Perform Arts.

Titze, I.R. and Durham, P.L.(1987), Passive Mechanism Influencing Fundamental Frequency Control, in *Laryngeal Function in Phonation and Respiration*, T. Baer, C. Sasaki and K. S. Harris(eds.), San Diego, College Hill Press, 304-319.

Umeda, N.(1975), Vowel Duration in American English, *J. Acoust. Soc. Am. 58*, 434-445.

Ushijima, T. and Sawashima, M.(1972), Fiberoptic Observation of Velar Movements during Speech, *Ann. Bull. U. Tokyo RILP 6*, 25-38.

Vaissiere, J.(1977), Quelques Experiences d'Analyse Perceptuelle en Français, *VIIIèmes Journées d'Étude sur la Parole, Aix-en-Provence, 25-27 Mai 1977*, 183-189.

Van den Berg, J.(1957), Sub-Glottal Pressure and Vibrations of the Vocal Folds, *Folia Phoniatrica 9*, 65-71.

Wood, S.(1979), A Radiographic Analysis of Constriction Locations for Vowels, *J. Phonetics 7*, 24-44.

Zealear, D.(1987), The Brainstem Connections with the Laryngeal Region of the Motor Cortex in the Monkey, in *Laryngeal Function in Phonation and Respiration*, T. Baer, C. Sasaki, K. S. Harris(eds.), San Diego, College Hill Press, 168-177.