MODERN METHODS OF INVESTIGATION IN SPEECH PRODUCTION

Osamu Fujimura, Bell Laboratories, Murray Hill, New Jersey 07974

Chairperson: Celia Scully

1. Descriptive Theory and Modeling of Speech Production

The process of speech production involves many aspects which may be treated by different disciplines of science. As much as we deal with speech as signals representing linguistic codes, it is clear that we need to have a descriptive framework of the linguistic message, so that we can relate the observed physical phenomena to the units that are used in the codes. Both segmental and suprasegmental specifications have to be given, as well as appropriate indications of surface syntactic (and semantic) information.

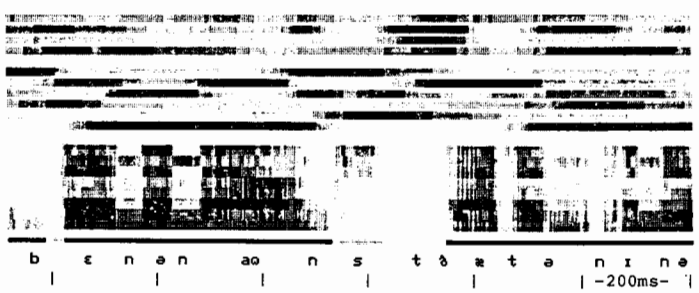
In addition to the lexically distinct accentual patterns and different intonational patterns for phrase structures, modulations of voice pitch and duration may be extensively used in conversational speech reflecting, e.g., focus, emphatic contrast, contextual and statistical predictabilities of the word, etc. Since speech phenomena always involve paralinguistic factors, such as the speaker's emotional state and idiosyncrasy, a way of describing those is also needed; or at least we must have a clear idea about what relevant factors have to be kept constant to make the comparison of different linguistic units meaningful. These considerations become more and more important, as we make progress in speech research. There are some emerging efforts in this direction, both in theory and experiment. The metric theory (Liberman and Prince 1977) for description of stress and intonation patterns of English constitutes a good example of such theoretical progress in this area, and a pitch contour synthesis-by-rule experiment based on this theory (Pierrehumbert 1979) suggests rapid progress in this field.

The notion of segments is also being revisited in connection with the significance of larger segmental units. The basic idea is to concatenate segmental units, whether phonemes, syllables, phonological words or phrases, to form larger units, and give suprasegmental modulations as patterns assigned to the larger units. Experiments in synthesis by rule attempt to evaluate models of this process. The notion of temporal modulation can be clarified only by referring to a well-defined model of speech dynamics that im-

plements an abstract specification of concatenated strings of units. Such a phonetic realization process would be characterized by different dynamic (i.e. temporal) characteristics for individual articulators, and the realized phonetic events corresponding to the so-called (phoneme size) "segments" are in general not in synchrony. Therefore, discontinuities observed in acoustic signals, such as the voice onset, stop release, etc., may not reveal some of the important aspects of the temporal characteristics of speech.

Gunnar Fant (1962) described a fine subsegmentation of acoustic signals based on their apparent discontinuities and interpreted such spectrographic representations of speech in terms of overlapping acoustic properties, roughly similar to, but crucially different from, the linguistic distinctive features.

FIGURE 1

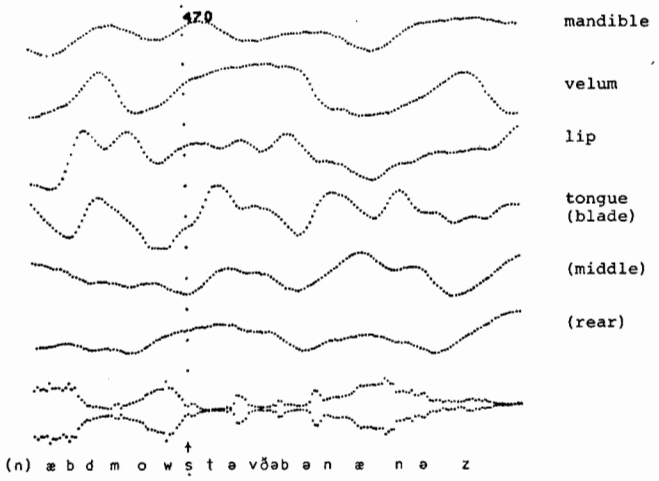


A combined articulatory-acoustic representation of part of a sentence 'Ben announced that an innocent-seeming infant had nimbly nabbed most of the bananas', uttered by a male native speaker of American English (Fresno, California). The upper part pertains to pellet positions, as obtained by the computer-controlled x-ray microbeam system, and the lower part a simplified (8 frequency-band) spectrographic pattern. In the lowest horizontal line black, gray, and white represent, respectively, voiced, voiceless, and silent states of the speech signal, and the phonetic symbols underneath are selected and placed automatically based on the articulatory information as well as the voicing state of the sound. The articulatory gesture is represented by the topmost 4 stripes for front (dark)/back (light) movements of the pellets placed on (from top) the lower lip, the blade, mid and rear portions of the tongue, and below these by the 6 stripes for up (dark) - down (light) movement of (from top) the lower lip, the mandible, the three parts of the tongue, and the velum (dark for low) (see Nelson [1979], Miller and Fujimura [1979]).

In order to account for the full information contained in speech signals and its human perception, one has to go well beyond this basic sketch. The spectral modulation of the speech signal is in one aspect discontinuous and in the other continuous. This dual nature of speech may be seen most obviously when we compare a gross spectrographic representation with an articulatory representation (see Figure 1) (Miller and Fujimura 1979). This qualitative difference betweeen articulatory movement and its consequent acoustic temporal pattern stems from the inherent non-linearity between the two levels of speech representation.

Recent studies are revealing interesting details of articulatory processes in relation to the phonological structures of the message. It is being shown that a simple model of concate-

FIGURE 2



Time functions representing vertical movements of the 6 pellets (the same material as in Fig. 1). The lowest trace depicts the speech waveform envelope. The arrow in line with a vertical array of dots is placed at the beginning of the voiceless segment for /st/.

nating phoneme-size units into larger phonological units, taking care of "coarticulation" phenomena by smoothing the movements, simply does not work. This is so particularly because within each syllable (or more exactly syllable core, see Fujimura and Lovins (1978), Fujimura (1979b)) there is something much more ad hoc about the temporal structure of phonetic events as syllabic ingredients. Such ad hoc characteristics are largely dependent on the language (and dialect) and therefore cannot be specified by a universal phonetic principle. By examining articulatory processes for relevant organs in movement, allowing for different dynamic characteristics and freedom of asynchrony in motor control for different articulatory (or phonatory) dimensions, we can obtain some insight into the nature of the temporal organization of phonetic events (Fujimura, forthcoming-a). Even inversions of temporal relations of peak activities for individual articulatory gestures are observed, from a phoneme string point of view. For example, as shown in Fig. 2, the syllable /mowst/ in a sentence utterance shows that the labial constriction for the glide /w/ manifests its peak activity during the voiceless period for /st/ toward the articulatory closure of /t/. A general principle governing phonetic structures of syllables (for the language) guarantees this looseness of temporal ordering within the syllable core to be irrelevant for phonological identification of this form (see Fujimura and Lovins (ibid)).

A useful descriptive framework thus seems to be one based on individual articulatory events related to elementary (functional) features of the syllable core. Basic notions, such as concatenation, coarticulation, assimilation and dissimilation have to be revisited quantitatively in light of such a descriptive model. It is time for us to produce experimental evidence for or against specific intuitive predictions. The scope of such experimental work is now being drastically expanded, thanks to newly available tools. It must be emphasized, however, that any of the available techniques for physical measurement, even in the future, is not likely to provide us with a complete picture of the physiologic/physical phenomena of speech production by itself. In order to interpret the results of measurements at different levels and relate them to each other, which is the task given to speech scientists for understanding the speech production process, we need to devise some new tools. Computational models of the natural speech production apparatus are being studied as such tools. For example, a three-dimensional static model of the tongue has been constructed using the finite element method (Kiritani et al. 1976) and is being used for studies of control characteristics of vowels (Fujimura and Kakita 1979).

A quantitative study of the gesture for the vowel [i], based on the tongue model, has suggested that the contraction of the posterior portion of the genioglossus muscle alone can give rise to a reasonable shape of the tongue and a consequent formant pattern for this vowel, but a slight deviation from the correct magnitude of contraction would cause quick deviation from the acceptable phonetic value. On the other hand, if we use a set of muscle components, in conformity with available electromyographic findings, we find such sensitivity to the degree of contraction is eliminated and the resultant phonetic quality becomes very stable and easy to achieve with a wide latitude of physiologic control. This points to the question of the quantal nature of speech as proposed by K. N. Stevens (1972), and also to the significance of feedback in different situations of speech production, including normal and artificial (such as the bite block) circumstances. With respect to the quantal nature, it seems that the crucial issue is the choice of the input level, at which the change of the controlled quantity in question is compared with that at the output, i.e. the acoustic characteristics such as formant patterns. The midsagittal tongue contour or a parametrically represented area function does not seem to be the correct input to the system for this specific discussion. The three-dimensional structure of the tongue combined with its volume incompressibility seems to play an essential role in characterizing the nonlinearity of the input-output mapping. Also, if, as our tongue model study seems to suggest, what is important in achieving a phonetic goal of articulatory gesture is selecting the pertinent set of muscles (with a certain balance of relative activities) rather than the exact magnitude of muscular contraction (excessive contraction resulting only in more or less unaffected physical consequences) the observed robustness of articulation under affected conditions seems more readily explicable than we had thought before. Gross orosensory feedback information also seems to play an important role in this connection (Perkell 1977).

Within the hierarchy of the natural process of speech production, the higher the level, the less applicable direct physical measurements are. Recent efforts by psychologists (see e.g. Sternberg et al., (1978)) are focused on temporal aspects of motor control, in the attempt to infer basic mechanisms of cortical programming and its execution. Studies of highly skilled performances in nonspeech areas seem to point to the understanding that in routine human actions the temporal course of a physical state takes a fixed preprogrammed pattern. In speech, articulatory events are decomposable into elementary gestures, such as lip movements for bilabial stops and velum raising for nasal-to-non-nasal transitions. Recent articulatory measurements indicate relatively constant speeds of such movements in a wide range of conditions when influences of certain separable factors are excluded (see the co-report on speech production by Sawashima (vol. I, p. 49-56)).

It has been argued (MacNeilage 1970) that the notion of invariant gestures for phonetic units is untenable in consideration of the high number of different contextual conditions. Such estimates, however, customarily depend on phoneme-size phonetic units as the basis of assuming targets. Based on an analysis that syllables are separable into cores and phonetic affixes, and each core into relatively constant dynamic patterns of initial and final demisyllables (the latter including the central portion of the syllable), we can actually construct for English a complete inventory of phonetic (concatenative) segmental units that contains less than 1,000 items for virtually all possible English phonetic forms (Lovins et al., 1979). Assuming that each inventory item is given phonetic indexes (syllable features) representing articulatory gestures, and also temporal parameters that are sensitive to nonsegmental conditions such as stress/accent, speed of utterance, etc., it does not seem implausible that the human brain can store all necessary phonetic patterns in the given language. An experimental evaluation of this new view is being attempted by synthesis-by-rule experiments using a demisyllabic inventory. A concrete model of acoustic realization of syllable features is being studied by Mattingly (1977). The psychological reality of the core-affix decomposition as well as the syllable itself is still to be examined.

2. Physiological Studies - Muscle Controls

The study of the physiology of speech production has seen remarkable progress in the past decade, even though there are still many unsolved basic questions. One general question is which muscle plays the principal role of implementing motor commands for a given phonetic gesture, viz. an elementary articulatory event. Electromyographic studies have revealed, for example, that the glottal abduction reflecting the devoicing gesture is related to the activity of the posterior cricoarytenoid muscles, whereas glottal adduction is achieved by several different muscles, including the interarytenoids, in varied ways depending on linguistic (and paralinguistic) functions (Hirose and Gay 1972; Hirose et al. 1978).

Hirano recently studied the anatomy and physiology of the vocal cords using various advanced techniques such as electron microscopy, histochemistry, electromyography, electric nerve stimulation, high speed motion picture, mechanical measurements, applied to both human and animal larynges (Hirano 1977). He arrived at an approximation of the complex anatomical structure by two (or three) loosely coupled parts, viz. cover and body. The cover seems to be responsible for the major part of the vibratory movement, showing large three-dimensional excursions, whereas the body contains the so-called vocalis muscle and participates in active parametric control of the vibrating system (see also Fujimura (1979a)). Baer (1975) has contributed a detailed study of excised canine larynges, and Titze and Talkin (1979) are contributing a new computerized model of the vocal cord vibration process.

Pitch control is an important topic from both lexical distinction and sentence-intonation points of view. The physiologic mechanism is not completely understood, but much is known now about the function of the cricothyroid muscle in relation to the voice fundamental frequency. There are cases where the voice fundamental frequency does not reflect the phonological accentual pattern because of the interaction between the consonantal control of voicing/tenseness and the vocal fold vibration frequency, but the electromyographic signal of the cricothyroid does (see Fujimura (forthcoming-b)).

Lingual muscles are difficult to study even with the best available electromyographic techniques because of the complex

interdigitation of a number of muscles forming the main body of the tongue. Nevertheless, the rather limited information obtained by EMG measurements are indispensable in inferring muscular functions relative to specific phonetic gestures.

Controlled interference by such techniques as anesthesia and bite block, has been experimentally induced, in order to evaluate the roles of feedback loops in speech production (Lindblom et al. 1977). In real utterance situations, mandible height is not necessarily correlated with tongue height either positively or negatively. For example, for the American English vowels /eɪ/ and /ɛ/ in sentence utterances, we have found in our X-ray microbeam data that a tongue height measure does distinguish occurrences of the two vowels very clearly, but that mandible height can be either lower or higher for one vowel than the other. Mandible height seems to reflect the stress status of the vowel, serving a function that is partially independent of the vowel height specification.

3. Physical States of Organs

Neural control of the larynx is parametric in the sense that gross average states of the larynx rather than details of vibratory changes of the peripheral shapes of the vocal cords are adjusted. For this reason, if we measure the laryngeal state during an utterance, the measurement may be taken at a relatively slow sampling rate such as 50 samples/second and averaged over a period like 20 msec. The fiberscopic technique developed at the University of Tokyo is appropriate for this purpose (Sawashima and Hirose 1968).

There have been successful studies of segmental control, such as manners of consonantal articulations in different languages (see for a review, Fujimura (1979a)). Here again, there are cases where the acoustic signal cannot answer a question about control. The laryngeal maneuver for pitch control seems related to vertical movements of the larynx as well as other gross appearances of the glottal area, and this may give us an opportunity to learn about pitch control even for devoiced syllables. A recent improvement of the fiberscope has made it possible to record two images side by side on the film stereoscopically, so we can measure the distance between the objective lens and the object (Fujimura et al. 1979). For many phonetic studies on qualitative states of the glottis, on the other hand, electric resistance measurements are being used as a readily applicable tool (Fourcin 1977, Frøkjær-Jensen 1968). Characteristics of voice source signals have gained renewed interest. Gunnar Fant (this volume, p. 79-108) is contributing a new insight about the interaction between the source and the vocal tract by closely examining speech waveforms. Flanagan et al. (1975) used their two-mass model of the vocal cords for simulating turbulence generation in the coupled source-vocal tract system.

The lips are obviously the easiest object to measure among different articulators, particularly with the use of a powerful stroboscopic technique (Fujimura 1961). A modern computerized system for measurement of the lips and mandible positions as well as linguapalatal contact is now available at the University of Alabama (McCutcheon et al. 1977). A servomechanistic technique can be used for a more general analysis of the natural articulatory systems such as the mandible and the lips. Such a measurement system has been implemented at the University of Wisconsin, Madison, and the control mechanisms of the lips are being studied assuming a linear system with feedback loops (Muller and Abbs 1979). The frequency response of such looped systems seems to allow actively controlled movements of visco-elastic systems via brainstem feedback for the majority of speech events. It should be emphasized, however, that the peripheral parts of articulators do not necessarily move together with the neurally controlled body of the same organ, and it is the former that determines acoustic consequences.

Dynamic characteristics of articulators in speech have been a vital issue in speech research. Several interesting proposals have been made about the basic principle of articulatory gestures trying to relate abstract and discrete phonological codes to the temporal structures of continuous speech phenomena (see Kent and Minifie (1977) for a review). Information on actual movements of the principal organs, in particular the tongue, is badly needed for such a study. Relatively large amounts of data obtained from the same subject are necessary to cope with an inherent variability of speech production phenomena. Collection of comparable data from many subjects, wherever possible, is another necessity for understanding the other aspect of human variability.

There are several methods that have been proposed and tested for observing tongue movements. Dynamic palatography (Fujimura et al. 1973b) represented an early attempt to computerize tongue observation for acquisition and processing of large amounts of data. It is also being applied to training of children in speech and hearing clinics in Japan. Other more recently proposed techniques include optical distance measurement between selected points on the palate and the nearest tongue surface. Magnetic (Sonoda 1977) as well as ultrasonic (Minifie et al. 1971) measurements also have been proposed.

The most direct and informative method of observing tongue movement is the use of X-rays for lateral views of the tongue. There used to be two factors that made radiographic measurements impractical for obtaining a large quantity of speech data. One is the radiological disturbance given to the subject. For this reason the exposure had to be limited usually to one or two minutes total per subject. The tedious and inefficient frame-by-frame analysis of the photographic images constituted another problem. The computer-controlled X-ray microbeam system was devised precisely to overcome these difficulties (Fujimura et al. 1973a). A full-scale system is now in operation at the University of Tokyo (Kiritani et al. 1975), and is producing useful results.

Several metal pellets are placed on selected points on the tongue and other articulators, usually but not necessarily in the midsagittal plane. A computer directs a thin X-ray beam to search around a predicted position, for each pellet, based on its past position and movement, verifies the current position, and repeats the procedure to look for the next pellet. By the combination of high sensitivity of the X-ray detector and an efficient use of the given total dosage for determining pellet positions, without exposing any unnecessary portions of the body for the specific purpose, the total radiographic exposure is incomparably smaller than that which would be used by film recording with an image-intensifier. The pellet position at each sample time, typically every 10 ms or less for 6-8 pellets, is digitally stored in the computer memory in real time. The experimenter, and the subject if desirable, can monitor the detected pellet movements. Powerful computer programs have been designed and implemented at Bell Laboratories in order to give the experimenter an efficient tool for interactive data analysis. Figure 1 represents one of the results, including an automatic annotation of the speech material with phonetic symbols (Nelson 1979).

An independent estimation of area functions by acoustic input impedance measurement has been proposed (Sondhi and Gopinath 1972). There is a nontrivial mapping process between the acoustically effective area function and the state of the speech organs (Mermelstein 1973). On the other hand, the so-called pseudo-area function that is conveniently derived by the well-established linear-prediction coding scheme (LPC) is not a true representation of the vocal tract characteristics proper (see Fant, p. 79-108, and Wakita, p. 151-172 (this volume)). Therefore, it is very desirable to have such independent measurement of the true area function, particularly if a simultaneous X-ray observation can be made for direct comparison of tongue shape (pellet positions) and the effective area function. The use of the recently developed CAT technique is also being attempted for static gestures.

4. Statistical Processing of Production Data

The availability of a large amount of production data encourages researchers to use advanced techniques of statistical processing of data such as multidimensional analysis (INDSCAL (Carroll and Chang, 1970) or PARAFAC (Harshman et al. 1974)), as well as principal component analyses. Through purely statistical processes, constituent (static) gesture components have been derived from both hand-traced midsagittal contours of the tongue of many speakers (Ladefoged 1977) and automatically tracked pellet position data for each of a few speakers (Kiritani and Imagawa 1976). These inductive methods give us purely phenomenologically derived "phonetic coordinates" for describing articulatory characteristics of a class of phonetic units, which is defined by the particular choice of the speech material used for this data processing. It is an intriguing question to ask if we can have a universal descriptive framework that explains the relations between different aspects of categorization of phonetic units (see Ladefoged's co-report on speech production (vol. I, p. 41-47)).

The use of multiple regression technique (both linear and nonlinear) must be mentioned in connection with the inverse mapping from acoustic characteristics to articulatory conditions. In addition to the more traditional method of analysis-by-synthesis,

which also is being used extensively (Fujisaki 1977), such new computational means seem to promise a new trend of research. Multiple regression techniques have been used for interpreting both durational parameters (Liberman 1978), and articulatory data (Nakajima 1977, Shirai and Honda 1977). The former used automatic processing of reiterant speech signals (Liberman and Streeter 1978), having the subjects mimic a sentence by a repetition of the same syllable, such as [ma], and attempted a best match between model-predicted and measured syllable durations by adjusting relative contributions of different phonologic and syntactic factors. The latter, using nonlinear regression, assumes a simple dynamic model of the physical movements of the articulators to determine the parameters that characterize such a physical system.

5. Concluding Remarks

When we define a domain of problems, such as normal speech, speech of a particular speaker, vowels as opposed to consonants, phonology as opposed to syntax, etc., we always need some understanding of the problems surrounding that domain. By knowing what happens just outside the boundary of the domain of immediate interest, in accordance with the principle of continuity, we always gain better insight as to how to delimit the domain. Thus, for example, speech pathology is another intriguing area of phonetic research. Needless to say, we would like to learn how people perceive speech, in order to investigate how people speak, because the real-life speech behavior is always a continuous mixture of production and perception.

References

Baer, T. (1975): Investigations of phonation using excised larynxes, PH.D. dissertation, M.I.T.

Carroll, J.D. and J. Chang (1970): "Analysis of individual differences in multidimentional scaling via an N-way generalization of 'Eckart-Young' decomposition", Psychometrika 35, 283-319.

Fant, G. (1962): "Descriptive analysis of the acoustic aspects of speech", Logos 5, 3-17.

Flanagan, J.L., K. Ishizaka, and K.L. Shipley (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell Syst. Tech. J. 54, 485-506.

Fourcin, A.J. and E. Abberton (1977): "Laryngograph studies of vocal-fold vibration", Phonetica 34, 313-315.

Frøkjær-Jensen, B. (1968): "Comparison between a Fabre glottograph and a photo-electric glottograph", Annual Report of the Institute of Phonetics, University of Copenhagen 3, 9-16.

Fujimura, O. (1961): "Bilabial stop and nasal consonants: A motion picture study and its acoustical implications", JSHR 4, 233-247.

Fujimura, O., S. Kiritani, and H. Ishida (1973a): "Computer controlled radiography for observation of movements of articulatory and other human organs", Comput. Biol. Med. 3, 371-384.

Fujimura, O., I.F. Tatsumi, and R. Kagaya (1973b): "Computational processing of palatographic patterns", JPh 1, 47-54.

Fujimura, O. and J. Lovins (1978): "Syllables as concatenative phonetic units", in Syllables and segments, A. Bell and J.B. Hooper (eds.), 107-120.

Fujimura, O. (1979a): "Physiological functions of the larynx in phonetic control", in Current issues in the phonetic sciences (Proc. of the IPS-77 Congress, Miami, Florida, Dec. 17-19, 1977) vol. I, 129-164, H. and P. Hollien (eds.), Amsterdam.

Fujimura, O. (1979b): "An analysis of English syllables as cores and affixes", Zs.f.Ph., Sign and system of language, Heft 4/5, 452-457.

Fujimura, O., T. Baer, and S. Niimi (1979): "A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation", JASA 65, 478-480.

Fujimura, O. and Y. Kakita (1979): "Remarks on quantitative description of the lingual articulation", in Frontiers of speech communication research, S. Öhman and B. Lindblom (eds.), 17-24, London: Academic Press.

Fujimura, O. (forthcoming-a): "Elementary gestures and temporal organization -- What does an articulatory constraint mean?", Proc. of the International Symposium on the Cognitive Representation of Speech in their series 'Advances in Psychology', G. Stelmach and P. Vroon (eds.).

Fujimura, O. (forthcoming-b): "Fiberoptic observation and measurement of vocal fold movement", Paper presented at the Conference on the Assessment of Vocal Pathology, NIH, Bethesda, Maryland, April 17-19.

Fujisaki, H. (1977): "Functional models of articulatory and phonatory dynamics", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 347-366, Tokyo: University of Tokyo Press.

Harshman, R., P. Ladefoged, L. Goldstein, and J. Declark (1974): "Factors underlying the articulatory and acoustic structure of vowels", JASA 55, 385.

Hirano, M. (1977): "Structure and vibratory behavior of the vocal folds", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 13-30, Tokyo: University of Tokyo Press.

Hirose, H. and T. Gay (1972): "The activity of the intrinsic laryngeal muscles in voicing control -- an electromyographic study", Phonetica 25, 140-164.

Hirose, H., H. Yoshioka, and S. Niimi (1978): "A cross language study of laryngeal adjustment in consonant production", University of Tokyo AB RILP 12, 61-72.

Kent, R.D. and D. Minifie (1977): "Coarticulation in recent speech production models", JPh 5, 115-133.

Kiritani, S., K. Itoh, and O. Fujimura (1975): "Tongue pellet tracking by a computer-controlled X-ray microbeam system", JASA 57, 1516-1520.

Kiritani, S. and H. Imagawa (1976): "Principal component analysis of tongue pellet movement", University of Tokyo AB RILP 10, 15-18.

Kiritani, S., F. Miyawaki, O. Fujimura, and J.E. Miller (1976): "A computational model of the tongue", University of Tokyo AB RILP 10, 243-251.

Kiritani, S., S. Sekimoto, and H. Imagawa (1977): "Parameter description of the tongue movements for vowels", University of Tokyo AB RILP 11, 31-38.

Ladefoged, P.N. (1977): "The description of tongue shapes", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 209-222, Tokyo: University of Tokyo Press.

Liberman, M.Y. (1978): "Modeling of duration patterns in reiterant speech", in Linguistic variation, models and methods, D. Sankoff (ed.), 127-138, New York: Academic Press.

Liberman, M.Y. and L.A. Streeter (1978): "Use of nonsense-syllable mimicry in the study of prosodic phenomena", JASA 63, 231-233.

Liberman, M.Y. and A. Prince (1977): "On stress and linguistic rhythm", Linguistic Inquiry 8, 249-336.

Lindblom, B. (1963): "Spectrographic study of vowel reduction", JASA 35, 1773-1781.

Lindblom, B., R. McAllister, and J. Lubker (1977): "Compensatory articulation and the modeling of normal speech production behavior", in Articulatory modeling and phonetics, R. Carré, R. Descout, and M. Wajskop (eds.), 148-161.

Lovins, J.B., M.J. Macchi, and O. Fujimura (1979): "A demisyllable inventory for speech synthesis", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 519-522.

MacNeilage, P. (1970): "The motor control of serial ordering of speech", Psychol. Rev. 77, 182-196.

Mattingly, I.G. (1977): "Syllable-based synthesis by rule", 9th International Congress on Acoustics, Madrid, July 4-9, 1977, Contributed papers 1, 512.

McCutcheon, M.J., S.G. Fletcher, and A. Hasegawa (1977): "Video-scanning system for measurement of lip and jaw motion", JASA 61, 1051-1055.

Mermelstein, P. (1973): "Articulatory model for the study of speech production", JASA 53, 1070-1082.

Miller, J.E. and O. Fujimura (1979): "A graphic display for combined presentation of acoustic and articulatory information", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 221-224.

Minifie, F.D., C.A. Kelsey, J.A. Zagzebski, and T.W. King (1971): "Ultrasonic scans of the dorsal surface of the tongue", JASA 49, 1857-1860.

Muller, E.M. and J.H. Abbs (1979): "Strain gauge transduction of lip and jaw motion in the midsagittal plane: refinement of a prototype system", JASA 65, 481-486.

Nakajima, T. (1977): "Identification of dynamic articulatory model by acoustic analysis", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 251-275, Tokyo: University of Tokyo Press.

Nelson, W.L. (1979): "Automatic alignment of phonetic transcriptions of continuous speech utterances with corresponding speech-articulation data", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 63-66.

Perkell, J.S. (1977): "Articulatory modeling, phonetic features and speech production strategies", in Articulatory modeling and phonetics, R. Carré, R. Descout, and M. Wajskop (eds.).

Pierrehumbert, J. (1979): "Intonation synthesis based on metrical grids", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 523-526.

Sawashima, M. (1979): "A supplementary report on speech production", Proc.Phon. 9, vol. I, 49-56.

Sawashima, M. and H. Hirose (1968): "New laryngoscopic technique by use of fiber optics", JASA 43, 168-169.

Shirai, K. and M. Honda (1977): "Estimation of articulatory motion", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 279-302, Tokyo: University of Tokyo Press.

Sondhi, M.M. and B. Gopinath (1972): "Determination of vocal-tract shape from impulse response at the lips", JASA 49, 1867-1873.

Sonoda, Y. (1977): "A high sensitivity magnetometer for measuring the tongue point movements", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 145-156, Tokyo: University of Tokyo Press.

Sternberg, S., S. Monsell, R.L. Knoll, and C.E. Wright (1978): Information processing in motor control and learning, G.E. Stelmach (ed.), 117-152, Academic Press.

Stevens, K.N. (1972): "The quantal nature of speech: evidence from articulatory-acoustic data", in Human communication, A unified view, P.B. Denes and E.E. David (eds.), 51-66, New York: McGraw-Hill.

Titze, I.R. and D.T. Talkin (1979): "A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation", JASA 66, 60-74.

DISCUSSION

H. Hirose, M. Hirano, and J.S. Perkell opened the discussion.

H. Hirose emphasized that we have to be careful in the interpretation of the electromyographic data, in particular because the relationship between the degree of muscle contraction and EMG output - in the case of speech muscles - is linear only under very special conditions. In order to get some idea of the relationship between the EMG pattern and the articulatory events we have to combine several methods. As an example, Hirose showed some EMG and X-ray microbeam data recorded simultaneously. He concluded that modern methods for the investigation of speech production can also be applied to the analysis of pathological patterns of movements and furthermore perhaps help towards a better understanding of the role of related parts of the central nervous system in speech production.

M. Hirano discussed various techniques employed for the study of the morphology and function of the vocal folds. He demonstrated that there are two different kinds of fibrous components in the vocal folds, namely the elastic and the collagenous fibres and showed how the vocal folds consist of more layers, partly from a histological and partly from a mechanical point of view (cf. vol. I, p. 189).

J.S. Perkell said that from his point of view the use of movement and EMG data along with sophisticated physiological modelling is only in its infancy with respect to the contribution that these techniques will eventually make to our understanding of speech production, dynamics, and hopefully also control strategies.

Then he commented upon the need for additional and better data in the third dimension, i.e. the cross-sectional area function, to supplement good midsagittal data. The need for such data is illustrated by the range of current notions that we have about factors, which underlie or constrain vowel categories. For example, B. Lindblom and his co-workers have proposed that a vowel category may be determined by an interaction between perceptual distance and some measure of ease of articulation; M. Lindau has proposed a primary role for the acoustic factors; K. Stevens has suggested some role for the patterns of tongue-to-maxilla contact; S. Wood and others have suggested that the vowel categories are determined by quasi-discontinuous relationships between the place of constriction and the sensitivity of formants to changes in place and degree of this constriction, along with factors related to the muscular anatomy; and Fujimura, working with the tongue model, has suggested a role for discontinuous relationships between muscle contractions and area function. Perkell noted that we have no way of disproving any of these hypotheses, and it may well be the case that to some extent all of them are valid. He concluded that to begin to untangle all the possible influences on vowel categories, we need a lot more well controlled work to test each one of these hypotheses, and that improved knowledge of area functions along with other factors is obviously essential for the evaluation of all the hypotheses on articulatory correlates of sound categories.

Then Perkell turned to the question about the X-ray dosage for different X-ray techniques. Perkell and his co-workers have made some dosage measurements and he gave the following values for the dosage that the subject would get:
10 rads/min for 35 mm conventional cineradiographic film (60 frames/sec); 2,5 rads/min for 35 mm high speed film and for 60 mm conventional cineradiographic film; 600 mrads/min for 16 mm high speed cineradiographic film; 260 mrads/min for video-tape.

Perkell noted that the microbeam system rarely gets above approximately half of these values, but under most circumstances the microbeam exposes the subject to a much smaller dosage. Finally, Perkell mentioned that the X-ray unit they are using allows for simultaneous views in the anterior-posterior and in the lateral dimensions, and he hoped that they might be able to obtain information which will contribute to our insufficient knowledge of area functions.

O. Fujimura confirmed that the dosage for the X-ray microbeam system is about one half of the smallest dosage obtained with other X-ray systems, namely 120 mrads/min. But the frame rate used for the estimate of 120 mrad is 120 frames per sec., i.e. twice the rate that Perkell used for his estimate. And in order to derive the total energy absorbed into the body, the dosage given should be multiplied by the area under exposure; since the 120 mrads estimated for the microbeam system assumes a constant exposure over a small area of 1 cm^2, the product is obviously 120, whereas the exposed area is much larger in the case of the two other X-ray systems.

E. Keller found the ultra sound technique a valuable alternative to the X-ray technique. The great advantage is that the exposure time can be considerably longer than the exposure time using cineradiography. Keller also pointed out that the frame rate is limited with the X-ray methods, which is a problem if we want to make measurements of speed of articulation, for instance. Finally, Keller said that with the ultrasonic method using a scanning beam, i.e. a system where a beam is sent back and forth several thousand times per sec., the whole surface of the tongue can be recorded, for instance, contrary to what can be obtained by a single beam system.

O. Fujimura claimed that for the X-ray microbeam system the net total of exposure time for one session is typically about 10 min., and often they run two or three sessions per subject. The total dosage given to the subject in terms of energy absorbed in one session is comparable to the amount of dosage one gets from the cosmic rays during one year. Concerning the limitation of frame rate, Fujimura replied that if one is interested in studying very fast movements in one portion of the tongue, which is the normal application of the ultrasonic technique, the number of pellets can be reduced and thereby a frame rate of up to 1000 frames per sec. can be obtained, so the frame rate for the microbeam system is not restricted to anything like 120 frames/sec.

Finally, Fujimura mentioned that for the velum height measurements - using the X-ray microbeam system - the pellet is not glued directly on to the velum as is the case with tongue pellets. Instead, a narrow strip of a very flexible plastic sheet is inserted through the nostril, covering the pellet, and this keeps the pellet in position, in contact with the upper surface of the velum.

J. Ohala emphasized that the estimates of radiographic dosage that we find in the literature vary tremendously. Furthermore, he referred to a study revealing an increased incidence of cancer in the thyroid from a population who had been radiated 30 years ago as children, with a dose of 6 rads, but these cancers did not develop until now. Ohala concluded that though the vocal tract is very important for us, we have to be very cautious in estimating our dosages. He advocated an intensification of our search for alternative ways of getting vocal tract informations.

G. Fant mentioned the possibility of measuring the impedance between two points, e.g. the upper and lower lip, as an alternative method for tracking the dynamics of articulation.

H. Künzel mentioned a very simple instrument for real-time recording of velar elevation, developed at the Institute of Phonetics in Kiel. The system consists of an optical probe - with an outer diameter of 3 mm - inserted through the nostril. The probe emits light which is reflected as a function of velar elevation. The linear function of the system has been controlled by simultaneous X-ray recordings.

C. Scully mentioned another approach, which works back from the aerodynamic stage and infers movements of the articulators from aerodynamic data. Such a technique can give us some idea of the size of the constrictor across which a pressure drop can be measured. What sort of range and what degree of accuracy this yields is an open question at the moment, but it is being investigated.

O. Fujimura mentioned a new technique, suggested by Dr. Sinada, where the pellet position is detected purely magnetically. The only disadvantage is that at the moment only one pellet can be tracked.

The indirect methods are very useful, in particular for practical purposes like clinical applications, training of articulatory gestures, and so on. But they need calibration and here the microbeam system could also be used.

S. Smith claimed that the electroglottographic method tells us something about the state of the musculature, i.e. whether it is relaxed or contracted.

O. Fujimura said that a technique for measuring the state of the muscular contraction by some physical means would be advantageous if we can establish a way to calibrate it.