SOURCES OF INTER- AND INTRA-SPEAKER VARIABILITY IN THE ACOUSTIC PROPERTIES OF SPEECH SOUNDS*

KENNETH N. STEVENS

Two practical problems related to speaker variability have attracted the attention of speech researchers in recent years. One of these is the identification or verification of a speaker from the speech sounds he produces. Research in this area attempts to answer the question: with what degree of reliability can an observer (through listening or through visual examination of spectrograms) or a machine identify a talker? The other application is determination of some aspects of a speaker's physiological or emotional state from measurements on his speech. The first of these applications requires that some attributes of an individual's speech remain fixed when producing sounds on different occasions, whereas the second examines the aspects of the speech that are susceptible to change from one occasion to the next.

Another reason for studying sources of speaker variability is to contribute toward a theory of speech production, perception, and acquisition. Such a theory should have two components: one component specifies the acoustic correlates of the linguistic units that are used for communication between speakers of a language, and the other describes the para-linguistic aspects of speech communication. The latter aspects are those which indicate the speech habits or characteristics of a particular talker, or which an individual uses to communicate particular emotions or emphasis.

In view of these theoretical and practical motivations, it is appropriate to examine the potential sources of inter- and intra-speaker variability, particularly as they are manifested in the acoustic properties of speech sounds. Our approach will be first to examine the mechanism of speech production and to indicate what aspects of this process are related to anatomical features that are likely to show differences from one individual to another, and what aspects are likely to be influenced by physiological and anatomical changes that can occur as a function of time for a given individual. After completing this review of the speech production process, we shall give several examples of acoustic data — some in terms of quantitative measurements and others in terms of qualitative observations — that illustrate these various sources of variability in the anatomy and physiology of speech.

For the most part, therefore, our concern is with aspects of speaker variability which can be explained in terms of anatomical differences or changes in physiological state, and not with those attributes that are learned by a speaker as a consequence of the linguistic environment in which his speech is acquired. There is not, however, a sharp dichotomy between aspects of speech production that are learned and those that are a consequence of anatomical or physiological attributes. The way a speaker learns to produce certain speech sounds may in fact depend on his particular anatomical characteristics and how they develop. It should be said at the outset that we are far from understanding the many causes of inter- and intra-speaker variability, and the examples given in this review can only serve to indicate the directions that future experimental and theoretical studies in this area might take.

1. SOME POTENTIAL SOURCES OF ANATOMICAL AND PHYSIOLOGICAL VARIABILITY IN SPEECH PRODUCTION

When he produces a given utterance, a talker presumably attempts to generate a sequence of sounds with specific acoustic attributes that enable a listener to decode the utterance into more or less the same linguistic representation as that of the talker. As we know, however, the sound wave for two utterances of the same word, phrase or sentence is never the same, whether these utterances are produced by the same speaker or by two different speakers.

In order to understand the kinds of anatomical and physiological changes that are likely to influence the speech of a talker, it is appropriate to review briefly the acoustic mechanism of speech production, and to indicate the articulatory structures that play primary roles in shaping the sound. As a part of this preliminary discussion, we shall speculate on the properties of these structures that are likely to differ from one individual to another, leading to inter-speaker differences in the attributes of the speech output, and the properties that can change with time for a given individual, leading to intra-speaker differences.

The conventional view of sound generation in the vocal tract is, of course, that of a sound source which is filtered by the vocal-tract resonators to produce a sound output from the mouth or nose (Fant 1960). For most speech sounds, the sound source results either from vocal-cord vibrations or from turbulent airflow at a constriction somewhere along the length of the vocal tract.

The structures that give rise to the sound sources and are responsible for shaping the acoustic cavities which filter these sources can, for our purposes, be divided into four classes, as shown in Figure 1. These are (1) the entire respiratory system below the larynx, including the airways (trachea, bronchi, etc.) and the lungs; (2) the larynx; (3) the vocal tract between the larynx and the lips, including both the pharyngeal and oral portions; and (4) the nasal cavity.

The dimensions and other properties of the subglottal respiratory system may vary greatly from one individual to another. The dimensions of the airways and the

* Preparation of this paper was supported in part by a grant from the United States Air Force and in part by a grant from the National Institutes of Health.

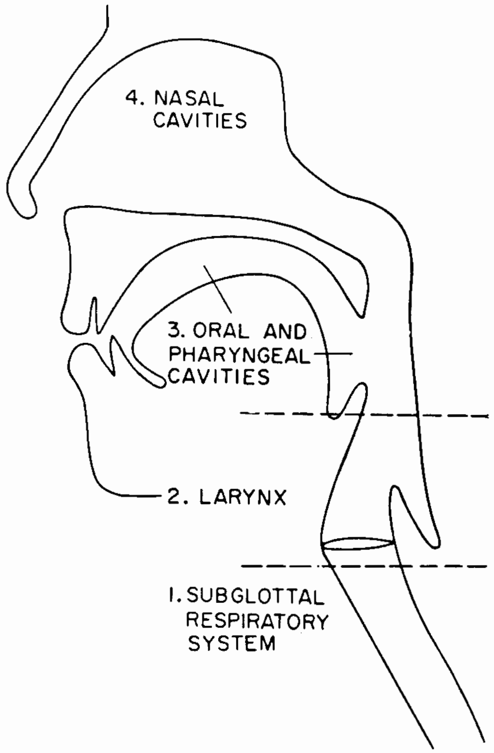
4. NASAL CAVITIES
3. ORAL AND PHARYNGEAL CAVITIES
2. LARYNX
1. SUBGLOTTAL RESPIRATORY SYSTEM

Fig. 1. Midsagittal section through vocal mechanism, identifying four regions that are potential sources of inter- and intra-speaker variability during speech production.

elasticity and size of the lungs may be different, and these result in differences in vital capacity and in the resistance of the subglottal airways. Some of these aspects may also vary from day to day for a given individual, and will certainly change over a longer period of several years. These changes will have only a secondary effect on the properties of the speech sounds, however, since the subglottal respiratory system is only indirectly involved in sound production.

The elasticity of lung tissue can have an indirect effect on speech, since it can influence the subglottal pressure and the rate at which the subglottal pressure can be varied. Since the subglottal pressure changes are responsible for some of the variation in fundamental frequency during speech, modification of the lung characteristics can result in changes in the contour of fundamental frequency and in the range of fundamental frequency used by an individual when he talks. Likewise, if a speaker is in some physiological state in which his respiration rate increases, his subglottal pressure during speech and hence his fundamental frequency are likely to increase.

The acoustic impedance looking down into the trachea from the glottis may be influenced by the configuration of the subglottal airways, and can have an effect on the pattern of vibration of the vocal cords. This impedance is responsible for the fluctuation in subglottal pressure during the glottal cycle, and the waveform of the glottal airflow will be determined in part by this pressure fluctuation. According to Lieberman (1967), it is also possible for this impedance to have an influence on the distribution of fundamental frequencies used by a speaker. The speaker tends to avoid using a fundamental frequency that is one-half of the lowest sub-glottal resonance (about 300 Hz for an adult male speaker).

The structure that is probably responsible for the greatest inter- and intra-speaker variability in speech is the larynx, and, in particular, the vocal cords. The waveform and frequency of the volume velocity through the glottis, which forms the source of acoustic excitation of the vocal tract for voiced sounds, is determined directly by the elasticity, mass and shape of the vocal cords. Any asymmetry in the vocal cords, such as a growth on one cord or a partial paralysis of one cord, can result in irregularities in the periodicity and waveform of the vibratory pattern. Abnormal dryness or increased wetness due to salivation can also lead to variations in surface tension and hence to changes in the vibration pattern. The fluctuating glottal opening causes modulations in formant bandwidth (at least for the first and possibly the second formants) and in formant frequency, and these effects will therefore depend on the glottal vibration pattern. Thus when we examine acoustic data for evidence of differences between speakers or differences within a speaker on different occasions or under various physiological conditions, we should direct our attention particularly to characteristics of the sounds that stem directly from the glottal source of vocal-tract excitation.

The third component of the speech production system — the vocal tract between the glottis and the lips — can be responsible for considerable inter-speaker variability, but probably does not change sufficiently from day to day to cause much intra-speaker variation. Since the vocal tract filters the glottal source for voiced sounds, differences in the dimensions of the vocal tract will give rise to different patterns of formant frequencies for a given vowel. Several ways in which the vocal tract of one individual might differ from that of another are illustrated in Figure 2. Examples of acoustic data corresponding to these kinds of vocal-tract differences will be given later.

Details of the vocal-tract shape, particularly in the region of the hard palate and the incisors, can influence the generation of fricative consonants, particularly those produced by raising the tongue blade. Both the intensity of the turbulence noise and the manner in which this noise is filtered by the cavities in the vicinity of the constriction can be affected by the configuration of the teeth and palate.

The nasal cavity plays a role in the production of nasal consonants and (in English) in the nasalized vowels that often occur adjacent to (usually preceding) nasal consonants. The size and configuration of the nasal cavity can differ appreciably from one individual to another. Furthermore, the membranes within the nasal cavity may expand or shrink so that the acoustic characteristics of this cavity vary with time for a given individual. The acoustic manifestation of these variations in the nasal cavity can best be observed in the spectrum of the nasal murmur during a nasal consonant, particularly the frequency and bandwidth of resonances that are determined primarily by the nasal cavity.

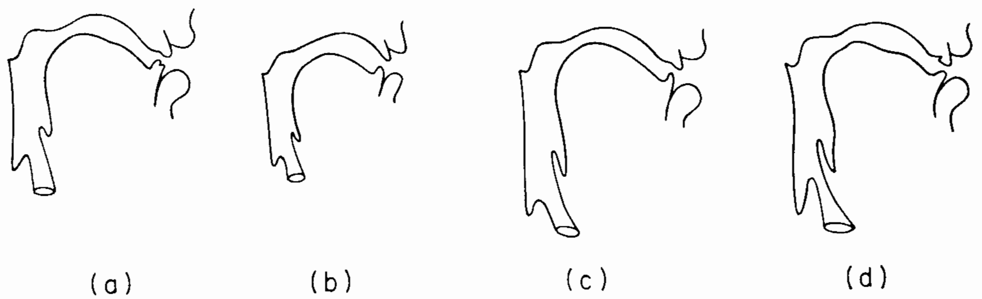

Fig. 2. Configurations (b), (c) and (d) indicate three ways in which vocal-tract dimensions might differ from those for configuration (a). In (b), all dimensions are scaled down in proportion, leading approximately to a proportional upward shift of all formant frequencies; in (c), the ratio of the length of the (vertical) pharyngeal portion to the oral portion is increased; in (d) there is a longer and narrower larynx tube, which could give rise to a relatively fixed higher-formant frequency (such as F4 or F5), since the resonance of this laiynx tube is largely uncoupled from the resonances of other portions of the vocal tract.

2. EXAMPLES OF ACOUSTIC DATA THAT DEMONSTRATE INTER- AND INTRA-SPEAKER VARIABILITY

1. *Fundamental Frequency of Glottal Vibration.* — Measurements of the fundamental frequency (F_0) and the range of fundamental frequency during speech have been made by a number of investigators. These studies include the measurement for different age-groups among children, young and middle-aged adults, and older adults. Systematic differences have been observed with increasing age in children up to 15-18 years. Beyond middle age, average fundamental frequency tends to increase with age (Mysak 1959). There are considerable differences in the fundamental frequency and in the distribution of fundamental frequency for individuals of the same sex and in the same age range. Examples of such distributions are shown in Figure 3 for six male college students, judged to be "superior speaker", reading factual prose (Fairbanks 1940). For some voices there appear to be ranges of F_0 that are used less often than higher or lower values of F_0 (Lieberman 1967), although this attribute is not evident for most of the data of Figure 3.

For a given individual, the average fundamental frequency, the fundamental-frequency range, and the shapes of the contours of fundamental frequency for a given utterance are strongly influenced by the emotional and physiological state of the talker. Results of a typical study are shown in figure 4 (Williams Stevens and Hecker 1970). This chart gives the average F_0 and range of F_0 (from 10th to 90th percentiles) for three different voices under several emotional situations. In general there is an increase in fundamental frequency for the emotions fear and anger, and a decrease for sorrow. These changes have been observed by a number of other investigators (Fairbanks 1940, Lieberman and Michaels 1962, Huttar 1968). The increases in fundamental frequency for the emotion fear and anger have been ascribed by Huttar (1968) to an increase in muscular tension throughout the body that is known to be a concomitant to emotion. Increased tension of the laryngeal muscles and the muscles controlling the respiratory system would tend to raise the fundamental frequency.

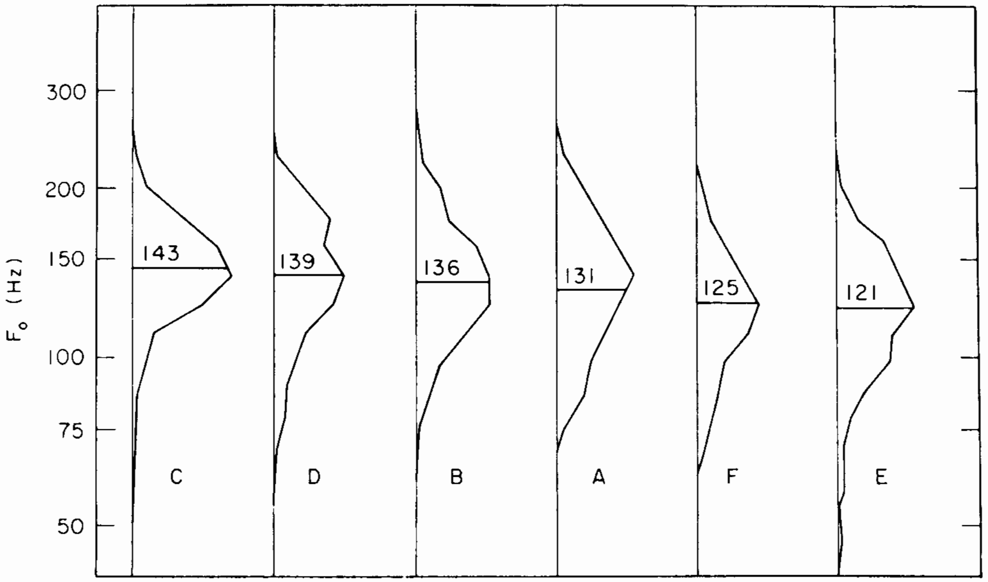

Fig. 3. Frequency distributions of fundamental frequencies used by six college-age superior speakers during the oral reading of factual prose. Medians are shown by the horizontal lines across the distributions (from Fairbanks 1940).

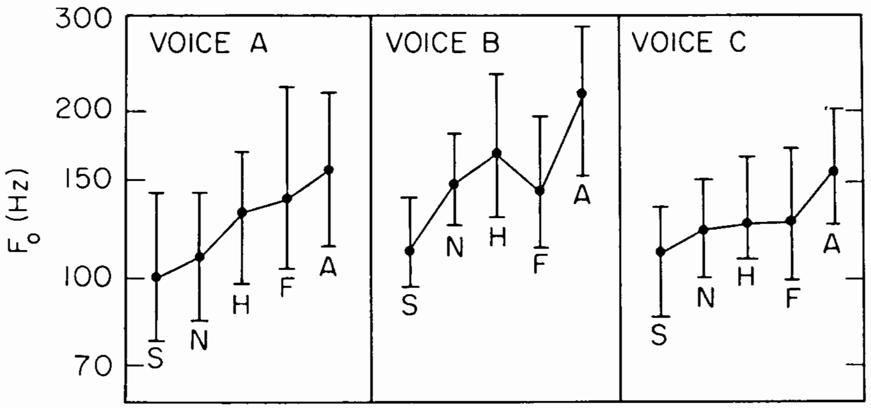

Fig. 4. Median fundamental frequencies and ranges (from 10 to 90 percent on cumulative distribution) for three actors producing several sentences in situations involving five different emotions: sorrow (S), neutral (N), happy (H), fear (F), anger (A).

For situations in which a talker experiences extreme fear, even larger increases in fundamental frequency have been observed (Williams and Stevens 1969). Task-induced stress has also been shown to cause changes in F_0, although the amount and direction of the change appears to vary from person to person (Hecker *et al.* 1968).

Typical contours of F_0 versus time for a speaker uttering the same sentence under different emotional conditions are given in figure 5 (Williams Stevens and Hecker 1970). For neutral utterances, the changes in F_0 are relatively slow, and the shape of the contour throughout each utterance is smooth and continuous. The contour shapes for utterances made in anger show a F_0 which is generally higher throughout the utterances, suggesting that they are generated with greater emphasis. Furthermore, either one or two syllables are characterized by clear peaks in F_0, again indicating strong emphasis on these syllables. Although the excursions in F_0 are quite great, there always appears to be a relatively smooth overall contour with one or two major peaks, but with no large discontinuities.

The contours for utterances made in situations involving the emotion sorrow are relatively flat with little fluctuations, and F_0 is usually lower than it is for a neutral situation. For this voice there is a slowly falling contour during the first half of the utterance, and a more level contour toward the end. Only rarely was emphasis placed on syllables in utterances produced by the three speakers in a sorrow situation.

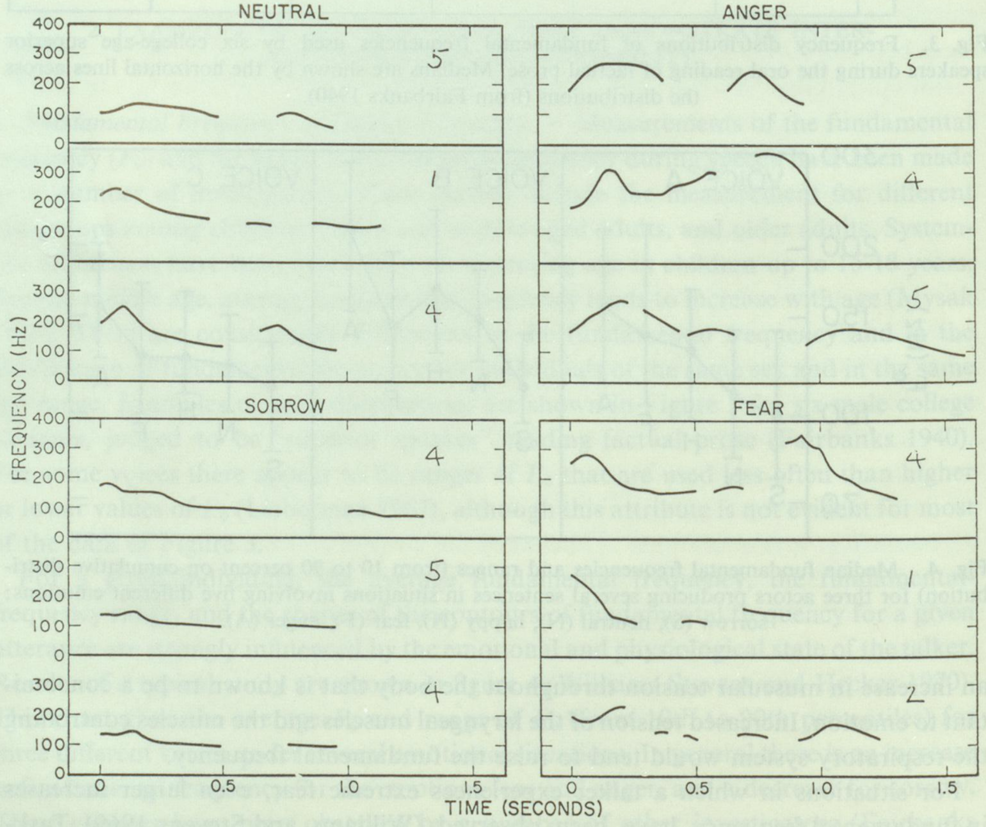

Fig. 5. Fundamental-frequency contours derived from utterances from one speaker under four different emotional situations. Each contour represents a word or a phrase (the phrases are identified by number).

For the emotion fear, the contours often depart from the prototype shape for the neutral condition. Occasionally there are rapid up-and-down fluctuations within a voiced interval, and sometimes sharp discontinuities occur from one syllable to the next.

In spite of these large variations in fundamental frequency for a given speaker, it has been found that data on fundamental frequency sampled at fixed points in a specified utterance spoken in a natural way by a cooperative speaker provide useful information to aid in verification of the speaker (Wolf 1969). On the other hand, since the F_0 contour is easy to modify or to mimic, and is strongly affected by the emotional state or stress of the talker, there are obviously situations in which measurements of F_0 cannot give a reliable indication of the identity of the speaker.

2. *Waveform of Glottal Volume Velocity.* — Evidence from spectral analysis of vowels shows that the waveform of the volume velocity through the glottis during a vibratory cycle can differ greatly from one speaker to another, and is presumably an important distinguishing characteristic of a speaker. Figure 6 compares spectrograms and spectra of the same vowel [i] produced by two speakers, where the different gross spectrum shapes must be ascribed to different glottal waveforms. The formant frequencies for the two speakers are approximately the same, but the amplitudes of

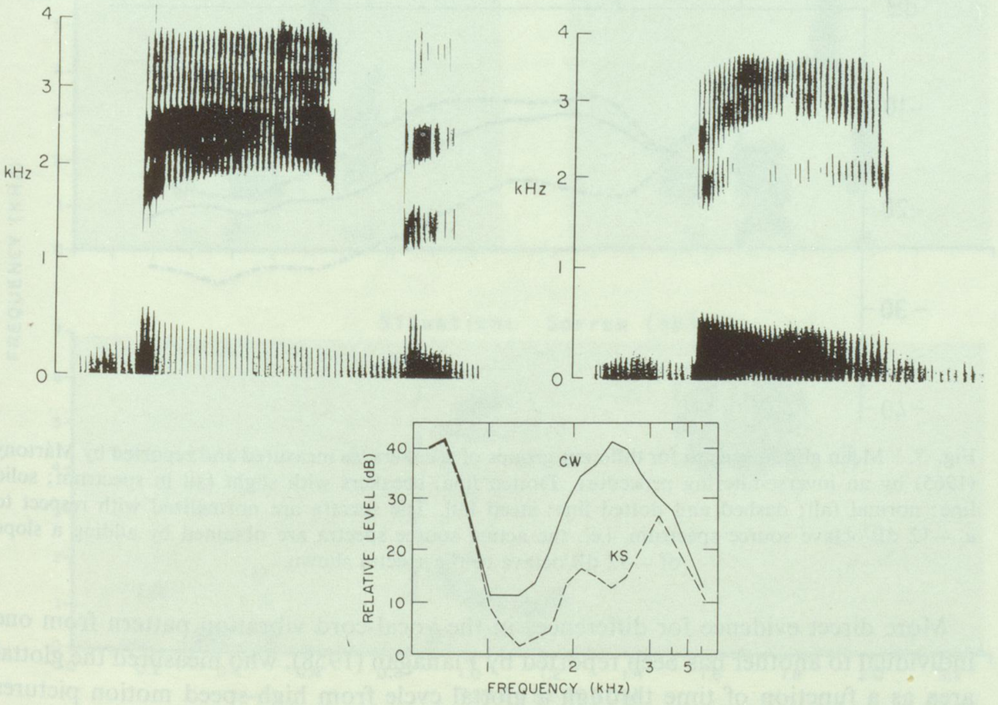

Fig. 6. Spectrograms of the word [bib] by two speakers. Also shown are spectra sampled in the vowel. The spectra were obtained from a bank of filters with bandwidths of about 360 Hz (or more for the filters centered above 2000 Hz). These data illustrate the large differences in glottal source spectra that can be observed for two normal speakers.

the third and higher formants relative to the amplitude of the first formant is much greater for one of the speakers, indicating a glottal pulse with more high-frequency energy. Presumably the pulse of glottal volume velocity for this speaker is narrower with a more abrupt discontinuity in slope during the closing phase of the cycle of vocal-cord vibration.

Mártony (1965) has measured the voice source spectrum for a number of speakers producing different vowels. He used a technique in which the formant frequencies were measured and the vocal-tract transfer function (in dB) was then calculated (together with the radiation characteristic) and subtracted from the vowel spectrum to yield an estimate of the source spectrum. Average spectra for three groups of his subjects are shown in Figure 7. Differences in amplitude in the third-formant region (about 2500 Hz) relative to that at low frequencies are as great as 10 dB. Mártony observed that for some speakers, the voice source spectrum depended to some extent on the vowel. Similar findings were reported by Carr and Trill (1964). Measures of the slope of the voice source spectrum were found by Wolf (1969) to provide information that was useful in distinguishing one speaker from another in a speaker identification scheme.

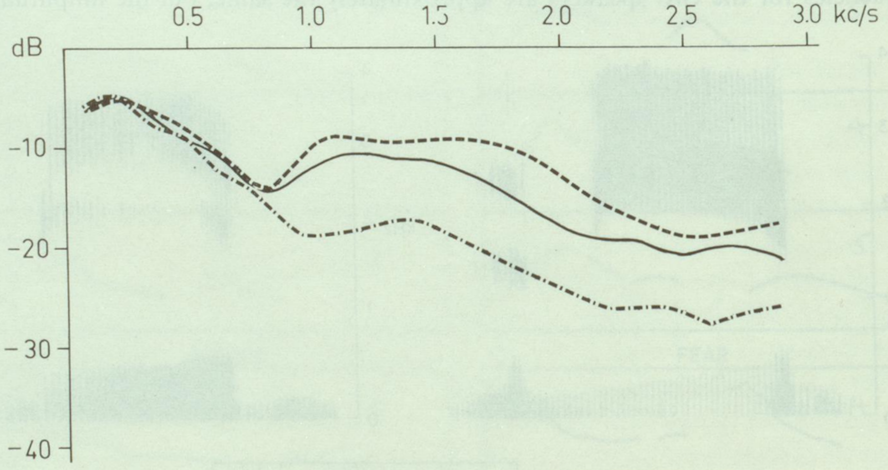

Fig. 7. Mean glottal spectra for different groups of speakers, as measured and reported by Mártony (1965) by an inverse-filtering procedure. Dotted line: speakers with slight fall in spectrum; solid line: normal fall; dashed and dotted line: steep fall. The spectra are normalized with respect to a —12 dB/octave source spectrum, i.e., the actual source spectra are obtained by adding a slope of —12 dB/octave to the spectra shown.

More direct evidence for differences in the vocal-cord vibration pattern from one individual to another has been reported by Flanagan (1958), who measured the glottal area as a function of time through a glottal cycle from high-speed motion pictures of the vocal cords.

The waveform of the glottal pulse is not fixed for a given individual but may change with the subglottal pressure and with the adjustment of the larynx musculature.

For example, several studies have shown that the glottal output is richer in high frequencies during speech at high vocal efforts. Abducting of the vocal cords can lead to 'breathy' voicing, whereas when the vocal cords are approximated, a condition of 'creaky' voicing can occur.

Spectrograms of voiced sounds often show irregularities in the amplitudes of successive glottal pulses, and these irregularities are usually most evident in the high-frequency region. A spectrogram illustrating this irregularity is shown in the spectrogram at the bottom of Figure 8. The lack of uniformity in the spectra of the glottal pulses is presumably due to the fact that successive pulses have slightly different shapes. This anomalous behavior may be the result of excessive turbulence in the glottal airflow, excessive moisture on the folds, or some asymmetry in the folds.

Irregularities of this type are sometimes an indication of hoarseness (Yanigahara 1967). They may also occur when an individual is under stress or is feeling strong emotion, particularly grief (Williams, Stevens and Hecker 1970). The spectrogram at the bottom of Figure 8, in fact, represents a sentence spoken by a talker who is expe-

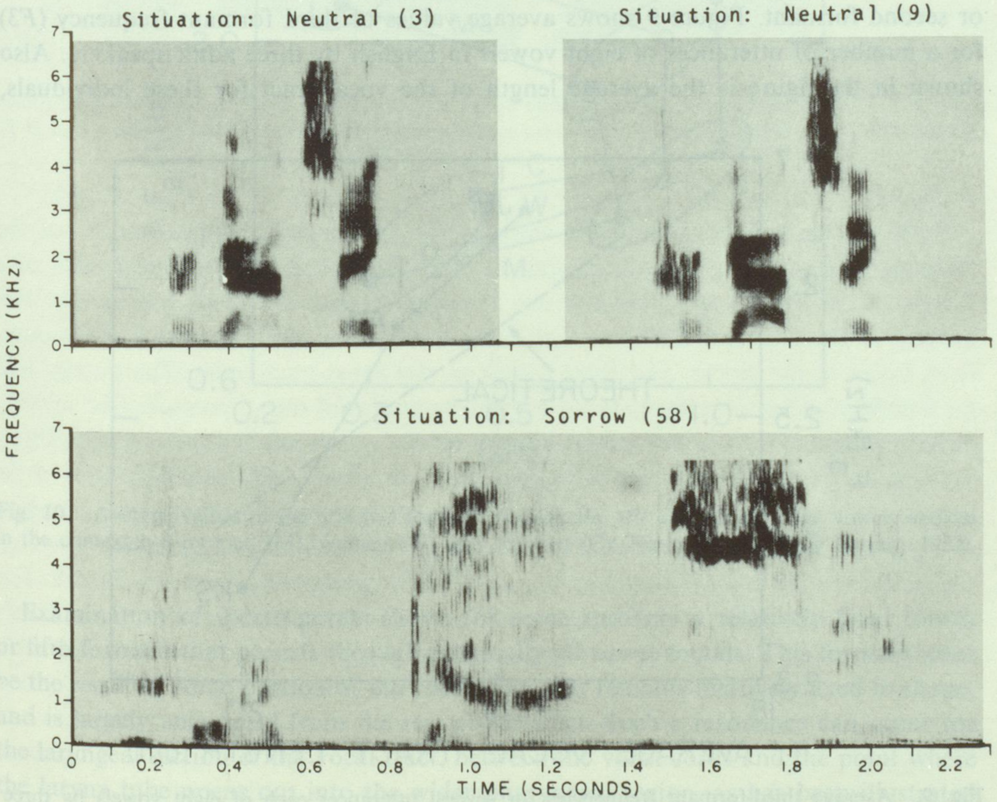

Fig. 8. Spectrograms of the same speaker saying the same sentence. The upper spectrograms were for a neutral emotional situation, and the lower one was in an emotional situation involving grief. Irregularity in the waveform of successive glottal pulses is very apparent during voiced portions of the lower utterance.

riencing grief. (Spectrograms of the same speaker producing the utterance in a neutral emotional situation are shown for comparison at the top of Figure 8.). Voicing irregularities may be characteristics of particular voices, reflecting some more-or-less permanent physical condition or pathology of the vocal cords (Lieberman 1963) or may be indicative of a particular physiological or emotional state for a talker, which may vary with time. Further work is needed to classify the various kinds of voicing irregularities that can occur.

3. *Formant Frequencies and Bandwidths.* — It is well known that there are differences in the formant frequencies for a given vowel produced by different speakers These variations may be due to differences in the dimensions of the vocal tract or in the learned speech habits of an individual.

If average values of formant frequencies are taken for a sufficiently large number of vowels, an indication of the average length of the vocal tract of a speaker is obtained. The average value of the third formant is particularly suitable for this purpose, since the third formant does not change markedly from vowel to vowel, and since it provides a more precise indication of average vocal-tract length than does the first or second formant. Figure 9 shows average values of third formant frequency (*F*3) for a number of utterances of eight vowels in English by three adult speakers. Also shown in the figuie is the average length of the vocal tract for these individuals, measured from X-ray pictures of several vowels (Stevens and House 1963). As the vocal-tract length increases, the formant frequency decreases, as theoretical considerations would predict. The dashed line indicates how the formant frequency would change if all the dimensions were scaled according to vocal-tract length; this line provides only a rough fit to the data.

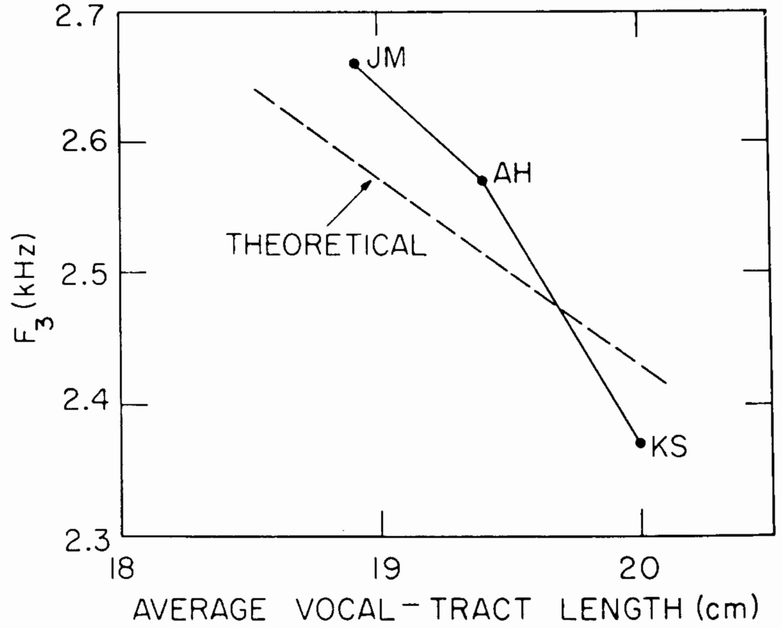

Fig. 9. Average third-formant frequencies for several utterances each of eight vowels by three different speakers, plotted as a function of the average vocal-tract length. The data show that the average formant frequency decreases as the vocal-tract length increases, although the amount of change does not exactly follow the predicted curve (dashed line), which is based on an assumption that all dimensions for vowel production are scaled in proportion to vocal-tract length.

More dramatic differences in formant frequencies are, of course, observed if data for men, women and children are compared, since the range of vocal tract lengths is much greater. Figure 10, for example, shows average values of the first two formant frequencies for several words for these three classes of speakers, as reported by Peterson and Barney (1952).

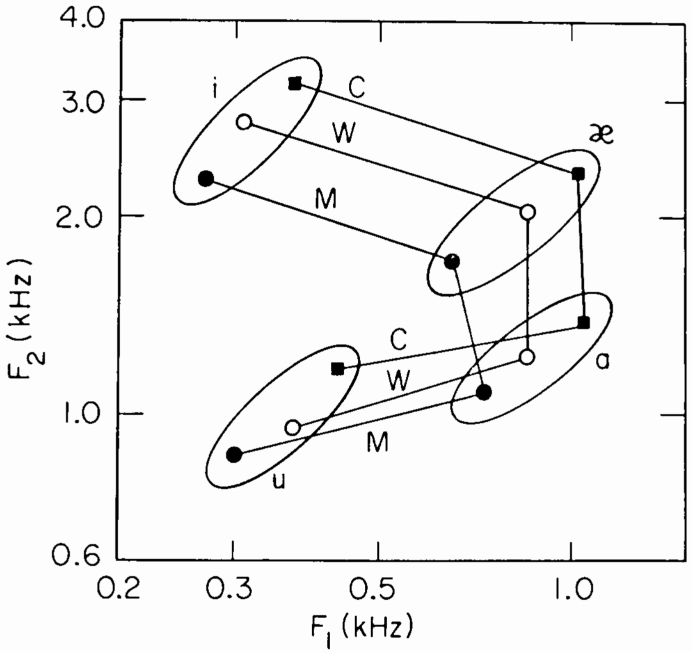

Fig. 10. Average values of the first two formant frequencies, (*F1* and *F2*) for four vowels spoken in the context h-d by men (M), women (W) and children (C) (From Peterson and Barney, 1952).

Examination of spectrograms shows for some speakers a relatively fixed fourth or fifth formant that persists through essentially all vowel sounds. This formant must be the result of some portion of the vocal tract that remains relatively fixed in shape, and is largely uncoupled from the rest of the tract. Such a resonance can occur for the laryngeal section of the vocal tract, between the vocal cords and the point where the larynx tube opens out into the wider pharyngeal region, as has been illustrated in Figure 2. Thus, for individuals having this anatomical characteristic of a relatively narrow larynx tube, a fixed high-frequency resonance, largely independent of vowel configuration, would be expected. A spectrogram of a sentence spoken by an individual with this characteristic is shown in Figure 11. The fixed resonance in this case is the fourth formant, and occurs at about 3000 Hz.

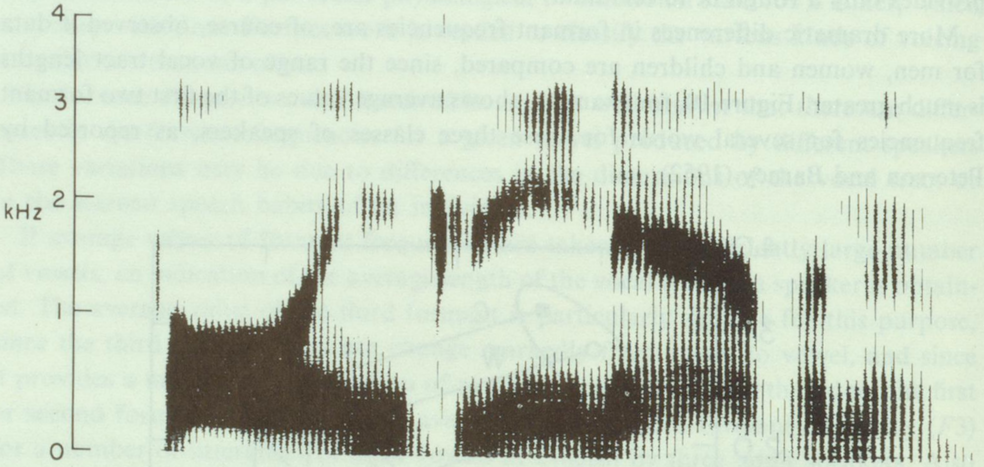

Fig. 11. A spectrogram of a phrase spoken by an individual who shows a relatively fixed fourth formant frequency independent of vocal-tract configuration during voiced sounds. The value of *F4* for this speaker is about 3000 Hz.

Perhaps less well documented is information concerning the bandwidths of the formants for particular vowels as spoken by different talkers. The bandwidth of a formant is proportional to the amount of energy loss in the vocal tract at the frequency of the formant. This loss may arise from sound radiation from the mouth, from losses at the glottis, and from losses at the vocal-tract walls. The energy loss for a particular formant depends upon the distribution of sound pressure and velocity in the vocal tract for that resonance. One vowel for which marked interspeaker differences in certain formant bandwidths have been observed is the vowel [i]. One of the formants for this vowel is the half-wavelength resonance of the narrow front portion of the vocal tract. Since this part of the tract is terminated in the mouth opening, the formant with this cavity affiliation would have a relatively large bandwidth. For some speakers, this is the third formant, and for others it is the second, probably depending in part on the relative lengths of the pharyngeal and oral cavities. Formant bandwidths for the vowel [i] for three speakers are given in Table 1. Two of these speakers (JM and AH) have a wide *F*3 and a narrow *F*2, whereas for KS the second formant bandwidth is relatively large. As expected, there are rather large differences in the shape of the spectrum envelope for these versions of the vowel [i] in the frequency range of the second, third, and fourth formants where the proximity of these formants leads to a broad spectral energy peak. Some examples of the spectrum shape for this vowel produced by several speakers are given in Figure 12. Large inter-speaker

TABLE 1

Bandwidths of first three formants (in Hz) of the vowel [i] for three different speakers. Averages are for six different utterances of the vowel (in various consonantal contexts) for each speaker. Note how speaker KS contrasts with the other two speakers.

	Speaker KS	Speaker JM	Speaker AH
B_1	40	60	40
B_2	220	60	60
B_3	110	320	200

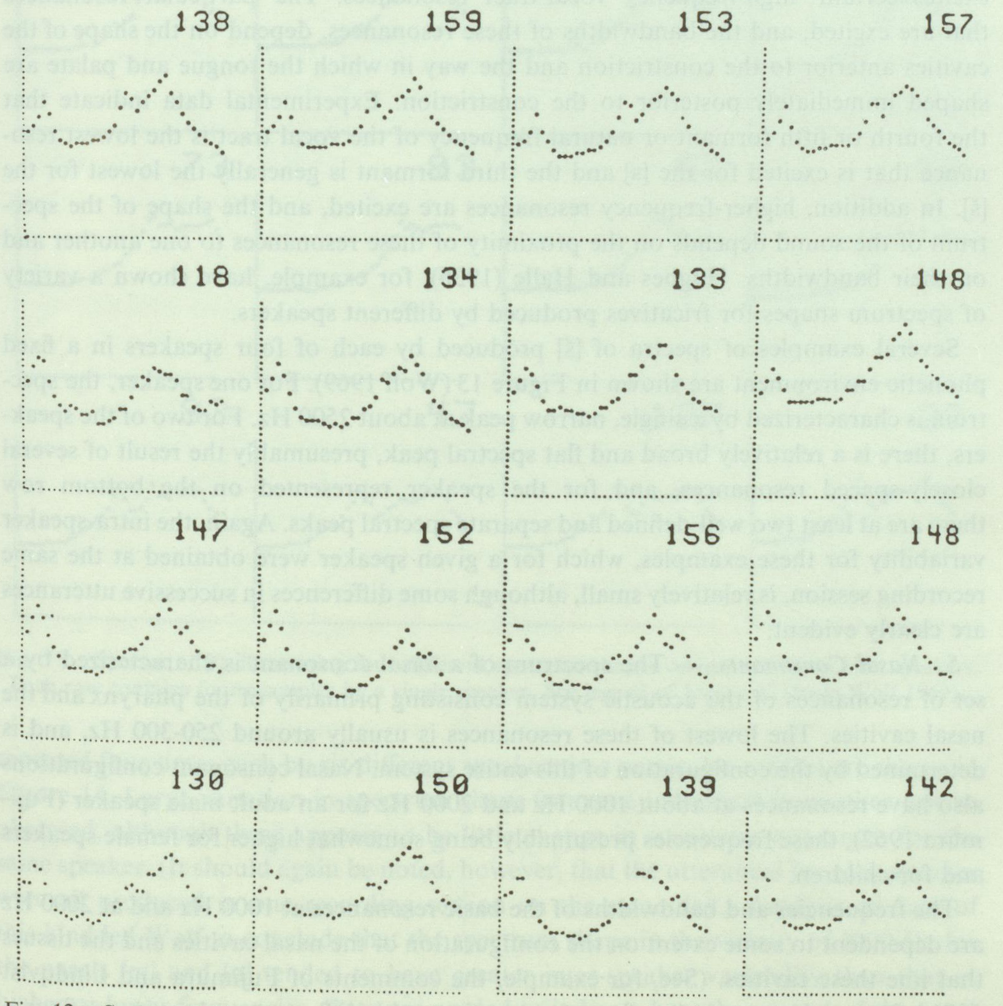

Fig. 12. Spectra of the vowel [i]. Each row contains four examples by a single speaker, and data for four different speakers are represented in different rows. The vertical scale is dB (2 dB per dot), and the horizontal frequency scale is from 150 to 7000 Hz (linear to 1600 Hz and logarithmic thereafter). Filter bandwidths are 150 Hz at low frequencies and increase to a maximum of 450 Hz at 7000 Hz. (From Wolf 1969.)

differences are apparent in these examples, whereas the differences for several utterances by the same speaker (all produced within a few minutes of each other on the same day) are relatively small. Wolf (1969) found that a measure of the shape of this spectral peak was useful in a speaker-recognition procedure based on acoustic measurements.

4. *Speech Sounds Produced with Turbulence Noise.* — The strident consonants [s] and [š] in English are produced by forming a narrow constriction between the tongue blade and the hard palate. Noise is generated by the turbulence that occurs when the rapid flow of air impinges on the alveolar ridge or the upper incisors, and this noise excites certain high-frequency vocal-tract resonances. The particular resonances that are excited, and the bandwidths of these resonances, depend on the shape of the cavities anterior to the constriction and the way in which the tongue and palate are shaped immediately posterior to the constriction. Experimental data indicate that the fourth or fifth formant or natural frequency of the vocal tract is the lowest resonance that is excited for the [s] and the third formant is generally the lowest for the [š]. In addition, higher-frequency resonances are excited, and the shape of the spectrum of the sound depends on the proximity of these resonances to one another and on their bandwidths. Hughes and Halle (1956), for example, have shown a variety of spectrum shapes for fricatives produced by different speakers.

Several examples of spectra of [š] produced by each of four speakers in a fixed phonetic environment are shown in Figure 13 (Wolf 1969). For one speaker, the spectrum is characterized by a single, narrow peak at about 2500 Hz. For two of the speakers, there is a relatively broad and flat spectral peak, presumably the result of several closely-spaced resonances, and for the speaker represented on the bottom row there are at least two well-defined and separate spectral peaks. Again, the intra-speaker variability for these examples, which for a given speaker were obtained at the same recording session, is relatively small, although some differences in successive utterances are clearly evident.

5. *Nasal Consonants.* — The spectrum of a nasal consonant is characterized by a set of resonances of the acoustic system consisting primarily of the pharynx and the nasal cavities. The lowest of these resonances is usually around 250-300 Hz, and is determined by the configuration of this entire system. Nasal consonant configurations also have resonances at about 1000 Hz and 2000 Hz for an adult male speaker (Fujimura 1962), these frequencies presumably being somewhat higher for female speakers and for children.

The frequencies and bandwidths of the basic resonances at 1000 Hz and at 2000 Hz are dependent to some extent on the configuration of the nasal cavities and the tissues that line these cavities. (See, for example, the comments of Fujimura and Lindqvist 1971.) One might expect, therefore, that speakers with different nasal cavity configurations would produce nasal consonants that differ in spectrum shape in the vicinity of 1000 and 2000 Hz. This variability in spectrum shape and its potential application to speaker recognition has been observed by Wolf (1969). Spectra of nasal murmurs

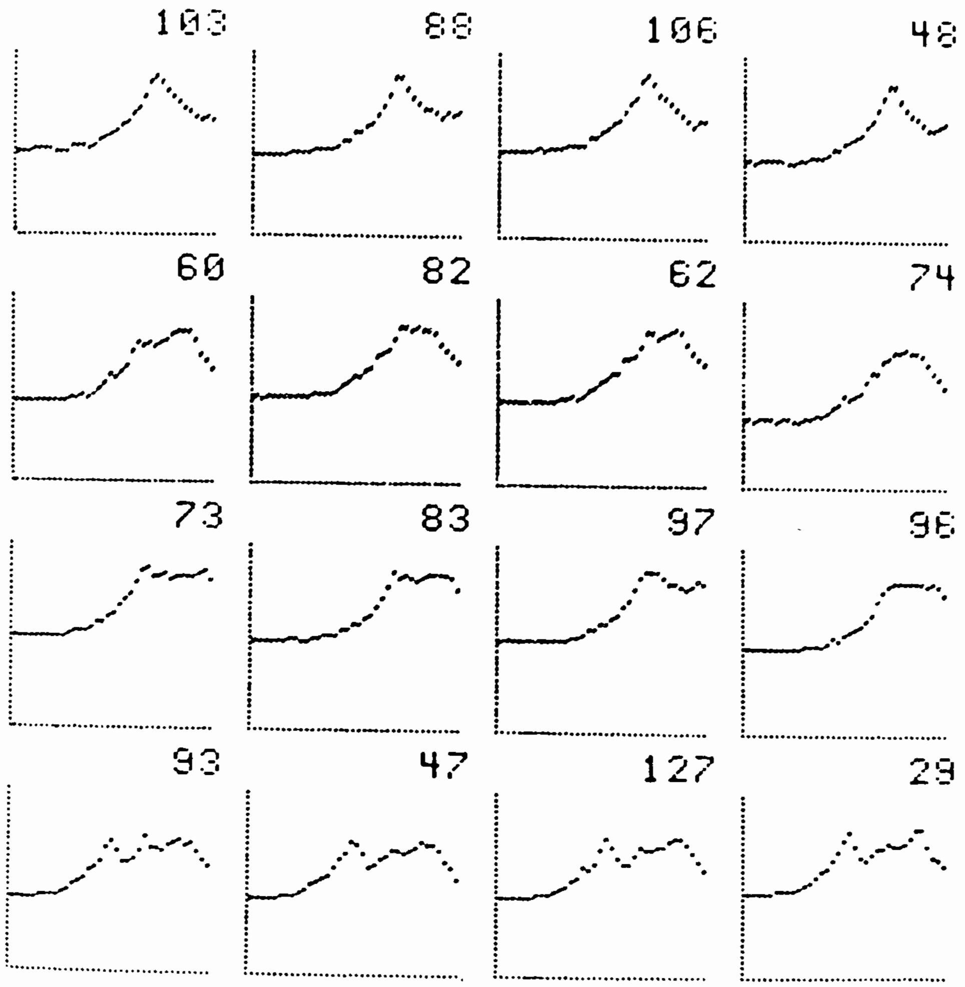

Fig. 13. Spectra of [š], illustrating four different spectrum shapes corresponding to four speakers. Each row contains four examples by a single speaker. See legend of Figure 12 (from Wolf 1969).

repeated four times each by six different speakers in a particular context are shown in Figure 14. Large variation in spectrum shape from one individual to another can be observed, although there appears to be little change in successive utterances for the same speaker. (It should again be noted, however, that the utterances for each speaker were all produced in one recording session on the same day.) Analysis of data of this kind led Wolf to conclude that the spectrum shape in the vicinity of 1000 Hz for the nasals [m] and [n] tended to have greater inter-speaker variability than that at higher or lower frequencies. Data are needed to indicate how the spectra of the nasal murmur for a given individual vary from day to day and week to week, since differences in the tissues within the nasal cavities are expected to occur from day to day.

6. *Some Other Sources of Acoustic Variability.* — Most of the acoustic charac-

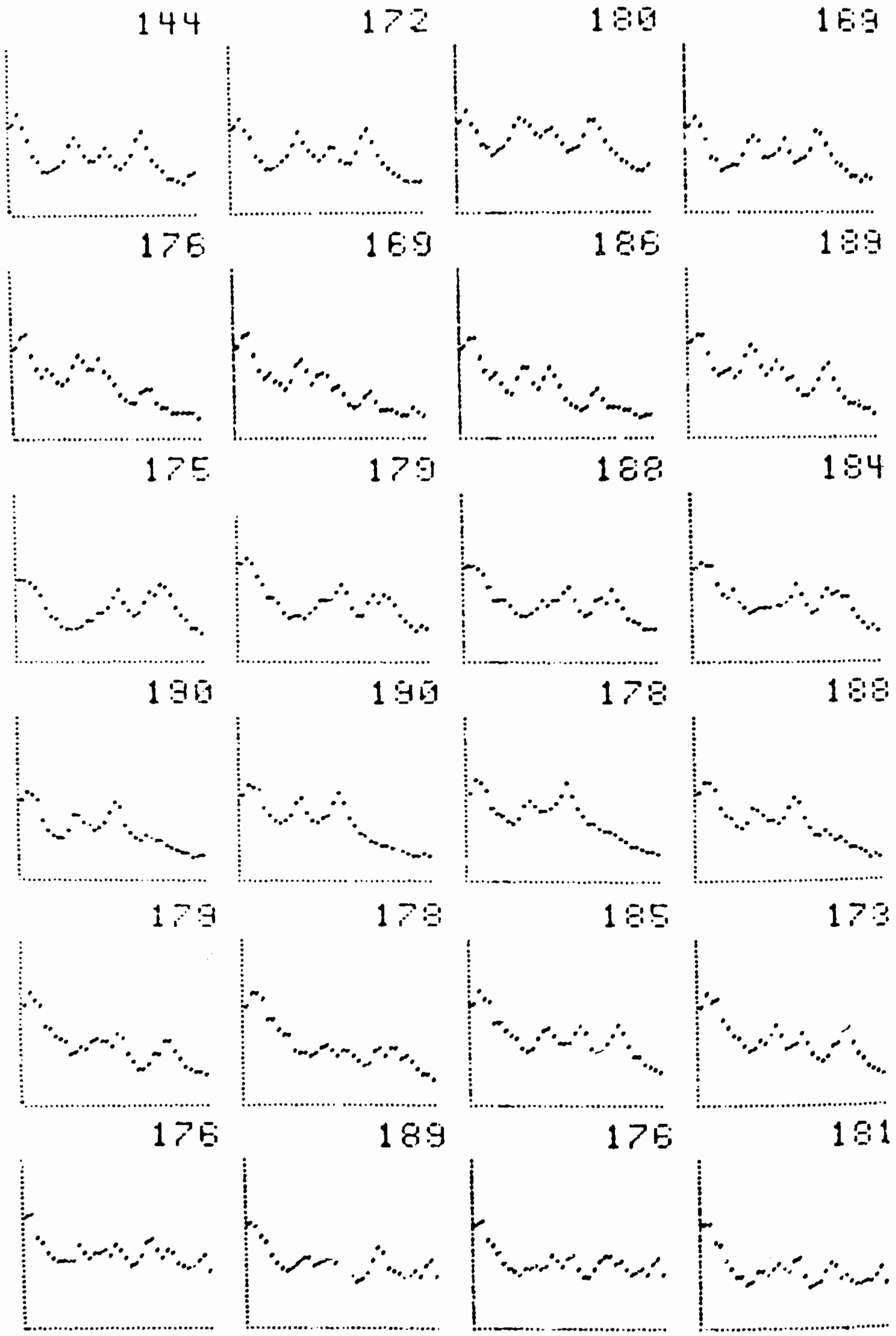

Fig. 14. Spectra of [m] in a fixed phonetic context. Each row contains four examples by one speaker, and different speakers are represented by different rows. See legend of Figure 12 (from Wolf 1969).

teristics that have been considered here are measured during steady-state regions of the sound output, when the vocal tract and larynx structures are in more-or-less steady configurations. It is to be expected, however, that differences or variability in articulatory dynamics account for a number of acoustic attributes that distinguish one speaker from another or that are likely to show intra-speaker variability. Inter-speaker variability in such attributes may be the result of learned articulatory habits rather than differences in anatomical or physiological characteristics of the articulatory structures, but nevertheless may often provide important cues for the identity of a talker. Those who have tried to perform a speaker identification from visual examination of spectrograms recognize the importance of patterns of formant movements, particularly during liquids, glides, and diphthongs. In their report of a series of experiments on speaker identification from spectrograms, Tosi *et al.* (1971) state that slopes of formants and durations are among the cues that subjects are instructed to use in the identification task.

Variability of speech sounds occuring in word, phrase and sentence material often arises because the context is sufficiently redundant that acoustic cues for all distinctive features are not required, or can undergo appreciable modification without impairing a listener's understanding of the utterance. Thus, a speaker has an option with regard to the precision with which he actualizes a feature that is predictable from the context; he may modify the actualization of a segment provided that this modification does not lead to a misinterpretation of the utterance. A few examples of these kinds of variability in English are: (1) prevoicing of an initial 'voiced' stop consonant is optional in English; (2) a final stop consonant may or may not be released in utterance final position; (3) various degrees of nasalization may occur for vowels that precede nasal consonants; (4) drastic modifications or omissions may occur in certain consonant sequences, particularly postdental consonants in the vicinity of unstressed vowels (e.g., secon*d* number, top o*f* the box, John'*s* shoe). Study of these sources of variability is of particular importance in the formulation of theories of the production and perception of sentence-type utterances. During rapid speech, what kinds of strategies do speakers follow in omitting or modifying particular acoustic cues when the context is known to be sufficiently redundant that those cues can be supplied by the listener? How is a listener able to make sense of this context to compensate for the lack of clear acoustic cues? To what extent are these modifications of the acoustic output speaker-dependent, and does a given speaker vary the strategy he uses to produce these distortions on different occasions?

3. APPLICATIONS

1. *Speaker Identification from Spectrograms.* — The application that has attracted the most interest, at least in the United States, is the identification of speakers from visual examination of spectrograms. The comments in the literature concerning the

effectiveness of this method of speaker identification have ranged from claims of almost error-free identification (Kersta 1962) to cautious statements that more research is necessary before the limitations of the method can be properly delineated (Bolt *et al.* 1969). The situation has been clarified somewhat with the publication of the results of an extensive study of voice identification from spectrograms by Tosi, *et al.* (1971). These investigators showed that, under the most ideal laboratory conditions, one speaker can be identified from a set of 10-40 speakers with an error of less than one percent. The ideal conditions include use of spectrograms of a number of clue words spoken in isolation; words to be matched were produced on the same occasion as the reference words, the spectrograms to be matched were produced by speakers who were known to have produced one set in the ensemble of reference spectrograms, and the observers had gone through a period of intensive training. At the other extreme, Tosi and his associates carried out experiments in which the clue words were the same for both matching and reference spectrograms, but they were produced in various contexts; the words to be matched and the reference words were produced on different occasions (separated by a month); and the spectrograms to be matched may or may not have been produced by speakers who were identified with the reference spectrograms. For these conditions, about 18 percent of the matches attempted by the subjects were incorrect. About 6 percent of the matches were false identifications, and 12 percent were false rejections.

Probably the most important reason for the difference in error scores for the two situations was the effect of using non-contemporary matching spectrograms, i.e., the error increased markedly when the matching spectrograms and the reference spectrograms were obtained from utterances recorded on different occasions a month apart. Apparently the intra-speaker variability was small for utterances produced within a few minutes of each other, but increased considerably when a month separated the sampling of the speakers' voices. The increase in error scores for non-contemporary matching spectrograms occurred both for false rejections and false identifications. These findings suggest that further study of this variable as it applies to the speaker identification task is needed. It has been shown, for example, that changes in emotional state or in the stress experienced by a speaker can result in changes in some attributes of his speech sounds. In the laboratory situation reported by Tosi *et al.* it can be assumed that the emotional or physiological state of the speakers in the two recording sessions one month apart was about the same, and hence the intra-speaker variability was held to a minimum. From this point of view, therefore, the Tosi results yield error scores that may be too small. On the other hand, an observer with sufficient training in interpretation of spectrograms may be able to identify the acoustic attributes that are least likely to be influenced by the state of the talker, and hence provide more reliable cues for the identity of the speaker. Tosi *et al.* emphasize the importance of training of the observers in the speaker identification task, and also notes that lower error scores are likely to be obtained if the scores are based only on those matches in which the observer has high confidence.

2. *Identification of Speakers by Automatic Procedures.* — A number of attempts have been made to devise procedures for the automatic identification of speakers through extraction of certain parameters or attributes from the acoustic signal. A review of this work is beyond the scope of this paper, since, with some exceptions, cited earlier, it has led to little insight concerning sources of speaker variability. Typically, proposed automatic speaker identification techniques have involved the recognition of a speaker from an ensemble of 10-20 unknown speakers. Error scores ranging from 0-10 percent have been obtained, but these scores appear to be based for the most part on the use of 'contemporary' utterances (in the same sense used by Tosi *et al.*).

In view of the wide variety of acoustic characteristics that contribute to inter-speaker variability, an optimum procedure for automatic identification of speaker should involve segmentation of the speech signal into regions characterized by various gross features before particular speaker-dependent details within these regions are identified. Thus, for example, detection of the presence of the vowel [i] is necessary before the details of the spectral peak associated with *F*2, *F*3 and *F*4 can be assessed. Or identification of nasals, voiced sounds, strident fricatives, etc. must precede the measurement of specific attributes that are characteristic of individual speakers.

3. *Monitoring the Emotional or Physiological State of a speaker from Measurement of his Speech Sounds.* — The emotional or physiological state of a speaker often has an effect on the properties of his speech, as indicated earlier in this paper, but at present most of these influences can be expressed only in qualitative terms. Furthermore, the acoustic effect of the emotional state may vary from one speaker to another. Probably the simplest quantitative measure that has the potential application to the monitoring of emotional state is the distribution of fundamental frequency. More research is needed, however, before the reliability of this or of other acoustic measures can be assessed (Hecker *et al.* 1968, Williams and Stevens 1969, Luk'yanov and Frolov 1969).

Department of Electrical Engineering and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Mass.

REFERENCES

Bolt, R.H., F.S. Cooper, E.E. David Jr., P.B. Denes, J.M. Pickett, and K.N. Stevens
1969 "Identification of a Speaker by Speech Spectrograms", *Science* 166:338-343.

Carr, P.B. and D. Trill
1964 "Long-Term Larynx Excitation Spectra", *Journal of the Acoustical Society of America* 36: 2033-2040.

Fairbanks, G.
1940 "Recent Experimental Investigations of Vocal Pitch in Speech", *Journal of the Acoustical Society of America* 11:457-466.

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

Flanagan, J.L.
1958 "Some Properties of the Glottal Sound Source", *Journal of Speech and Hearing Research* 1:99-116.

Fujimura, O.
1962 "Analysis of Nasal Consonants", *Journal of the Acoustical Society of America* 34:1865-1875

Fujimura, O. and J. Lindqvist
1971 "Sweep Tone Measurements of Vocal-Tract Characteristics", *Journal of the Acoustical Society of America* 49:541-558.

Glenn, J.W. and N. Kleiner
1968 "Speaker Identification Based on Nasal Phonation", *Journal of the Acoustical Society of America* 43:368-372.

Hecker, M.H.L., K.N. Stevens, G. von Bismarck, and C.E. Williams
1968 "Manifestations of Task-Induced Stress in the Acoustic Speech Signal", *Journal of the Acoustical Society of America* 44:993-1001.

Hughes, G.W. and M. Halle
1956 "Spectral Properties of Fricative Consonants", *Journal of the Acoustical Society of America* 28:303-310.

Huttar, G.L.
1968 "Relations Between Prosodic Variables and Emotions in Normal American English Utterances", *Journal of Speech and Hearing Research* 11:481-487.

Kersta, L.
1962 "Voiceprint Identification", *Nature* 196:1253-1257.

Lieberman, P.
1963 "Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathological Larynges", *Journal of the Acoustical Society of America* 35:344-353.
1967 *Intonation, Perception and Language* (Cambridge, Mass., M.I.T. Press).

Lieberman, P., and S.B. Michaels
1962 "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech", *Journal of the Acoustical Society of America* 34:922-927.

Luk'yanov, A.N. and M.V. Frolov
1969 *Signals of Human Operator State* (Moscow, Nauka Press) [NASA Technical Translation F-609].

Mártony, J.
1965 "Studies of the Voice Source", *STL-QPSR*:4-9 (Royal Institute of Technology, Stockholm, Sweden).

Mysak, E.
1959 "Pitch and Duration Characteristics of Older Males", *Journal of Speech and Hearing Research* 2:46-54.

Peterson G.E. and H.L., Barney,
1952 "Control Methods Used in a Study of Vowels", *Journal of the Acoustical Society of America* 24:175-184.

Stevens, K.N. and A.S. House
1963 "Perturbation of Vowel Articulations by Consonantal Context: an Acoustical Study", *Journal of Speech and Hearing Research* 6:111-128.

Tosi, O., H.J. Oyer, W.B. Lashbrook, C. Pedrey, and J. Nichol
1971 "Voice Identification Through Acoustic Spectrography", Report issued by Department of Audiology and Speech Sciences, Michigan State University (East Lansing, Michigan) February.

Williams, C.E. and K.N. Stevens
1969 "On Determining the Emotional State of Pilots During Flight: an Exploratory Study", *Aerospace Medicine* 40:1369-1372.

Williams, C.E., K.N. Stevens, and M.H.L. Hecker
1970 "Acoustical Manifestations of Emotional Speech", *Journal of the Acoustical Society of America* 47:66(A).

Wolf, J.J.
1969 "Acoustic Measurements for Speaker Recognition", unpublished PhD. thesis (Massachusetts Institute of Technology).

Yanagihara, N.
1967 "Hoarseness, Investigation of the Physiological Mechanisms", *Annals of Otology, Rhinology and Laryngology* 76:472-488.

DISCUSSION

EMOTO (Fukuoka)

First of all, I should like to express my hearty thanks to you, Dr. Stevens, for your report on your research concerning sources of inter- and intra-speaker variability in the ACOUSTIC properties of speech sounds.

To my mind, as Dr. Stevens also mentions, strictly speaking, no two sounds are exactly alike, because sounds are produced with different energies each time by the same speaker or by two different speakers in a sequence of time in an ever-changing situation. In other words, the speaker's or the hearer's psychological and physiological or physical conditions in relation to his environmental circumstances are the factors that determine the changes. Viewed in this light, there can be no such things as are called identical speech-sounds or phones to be reproduced, or phonemes to be realized or allophones to be considered the realizations of phonemes. All that comes into being will be so-called sounds — each different from the other.

However, there may be some features or attributes of SOUNDS UTTERED that are considered to be fairly similar to each other, which could be classified in some groups, depending on degrees of abstraction or perception. Namely, some such sound features or attributes should be arranged according to the similarities of those sounds uttered — similarities viewed from all angles such as functional, distributional, significant, phonetic, phonemic, allophonic, phonological, prosodic, physical, physiological, acoustic, vocal, phenomenological, psychological, or whatever you may call it, properties. These groups will perhaps turn out to be some keys or clues to the identification of a talker or a speaker to some degree, although the ACOUSTIC correlates of the LINGUISTIC units that are used for communication between speakers of a language, if it means a kind of one and the same idiolect, are always dependent on the hearer's physio-psychological conditions.

Since, in my view, the acquisition and the production of speech sounds of a particular language, which must go through the brain anyway, are largely dependent on habit-formation, and habit being a sort of second nature, some attributes of an individual's speech may be considered to remain fixed when producing sounds on different occasions, as Dr. Stevens states. So, some such speech habits, that is, sounds produced by, or a kind of voice prints of, a particular talker, may be considered fairly fixed, and it may be said that some attributes or aspects of such speech habits will be the natural consequence of the physical or physiological conditions or state of the speech organs of the particular talker.

As far as SPEECH perception is concerned, since meaning is involved, it will not be always easy to determine the aspects that indicate the speech habits or characteristics of a particular talker, or which an individual uses to communicate particular emotions or emphasis, unless the aspects of his speech habits or characteristics are closely studied in relation to the meaning he wants to convey by them; because in SPEECH communication there is such a thing as is called misunderstanding or misinterpretation.

For instance, Leonard Bloomfield says, "The meaning of a linguistic form is the situation in which the speaker utters it and the response which it calls forth in the hearer". In the dialogue between *A* and *B* such as, "What did you see in the zoo yesterday?" — "We saw two *liars* and tigers", one might think that *B* said, "We saw two *lions* and tigers". And such will often be the case in actual conversation. According to Prof. A.C. Gimson of University College London,

> It is well to remember that, although the sound system of our spoken language serves us primarily as a medium of communication, its efficiency as such as an instrument of communication does not depend upon the perfect production and reception of every single element of speech. A speaker will, in almost any utterance, provide the listener with far more cues than he needs for easy communication. In the first place, the situation, or context, will itself delimit very largely the purport of an utterance. Thus, in any discussion about a zoo, involving a statement such as 'We saw the lions and tigers', we are predisposed by the context to understand 'lions', even though the *n* is omitted and the word actually said is 'liars'. Or again, we are conditioned by grammatical probabilities, so that a particular sound may lose much of its significance.

and so on. In extreme cases, as Prof. David Abercrombie of the University of Edinburgh writes,

> The actual sense of the words used in phatic communion matters little; it is facial expression and intonation that are probably the important things. It is said that Dorothy Parker, alone and rather bored at a party, was asked, 'How are you? What have you been doing?' by a succession of distant acquaintances. To each she replied, 'I've just killed my husband with an axe, and I feel fine'. Her intonation and expression were appropriate to party small talk, and with a smile and a nod each acquaintance, unastonished, drifted on.

So, Dr. Stevens will be right in saying that "our concern is with aspects of speaker variability which can be explained in terms of anatomical differences or changes in physiological state".

However, it will have to be borne in mind that speech sounds are identified or judged as such by the listener with his memory or knowledge or sound image as the clue. In other words, an actual sound is determined and classified as such with the memory of some abstract sounds or abstracted sound features as clues. Therefore, the determination, or classification, or recognition of sounds, is largely dependent upon one's way or degree of perception or abstraction or generalization or LINGUISTIC analysis of the speech uttered. So, speech may be most objectively and scientifically studied by well-trained phoneticians and linguists through their objective and accurate observations and judgements based on their subjectivity, and with precision machines as aids.

At any rate, it is interesting to learn from Dr. Stevens that when we examine acoustic data for evidence of differences between speakers, or differences within a speaker on different occasions or under various physiological conditions, we should direct our attention particularly to characteristics of the sounds that stem directly from the glottal source of vocal-tract excitation, and so forth. Dr. Stevens has further presented to us a detailed report of his research on how an observer (through listening or through visual examination of spectrograms), or a machine, can identify a talker and how some aspects of a speaker's physiological or emotional state can be determined from measurements on his speech. I thank you again, Dr. Stevens, for your valuable report.

FRANCESCATO (Amsterdam)

I wish to add something to the remarks already made by Prof. Emoto on the paper by Dr. Stevens. One point I missed in his report is reference to the socio-linguistic aspects of phonetic variation. When dealing with speaker identification through language we cannot forget that speech variability is strictly linked with the extra linguistic features of that situation and of the reaction of the speaker to them. Labels of INTER- and INTRA-SPEAKER VARIATION are not enough, it seems to me, to take care of this condition. I think we are entitled to distinguish between two plans of variation, one of INHERENT (or PERMANENT) variation and one of SITUATION-CONDITIONED variation, so that we may be able to think in this connection not only of SPEAKER identification but also of SITUATION identification.

STEVENS

Dr. Emoto's comments emphasize the role played by context in determining the acoustic form of an utterance and in determining how the utterance is perceived by a listener. Since the speech signal provides us with more cues than are required, there is considerable opportunity for variability to occur in these acoustic cues. Different contextual and social situations can lead to differences in the way speech sounds, words, and phrases are produced. Dr. Emoto's remarks on these matters are most pertinent, as are those of Dr. Francescato.

LINDQVIST (Stockholm)

There is not much to add to this well-written paper which has drawn attention to the mechanisms underlying the variability in speech production. The importance of knowledge about these facts has also been emphazised by recent interests in making hi-fidelity speech synthesis to be used as a research tool. Fant and Ohman have made a few experiments about what aspects of the speech is possible to mimic. These experiments indicate that the dynamics of the voice source is most easy to imitate.

When we identify a speaker by listening to his voice we mainly use these features, which are intonation, dynamics of the source spectrum and timing. This means that a machine that should recognize a speaker cannot use these features.

I would like to ask two questions.

Many speakers have a minimum around 1 kHz in their voice source. It has been shown by van den Berg's experiments and by sweep-tone measurements of the vocal tract transfer function that this minimum is caused by absorption of energy down into the trachea. The frequency should accordingly be stable in frequency and it seems to be the only feature that the subject cannot change by will. I would like to ask if there have been any experiments made to use this as a cue for speaker identification.

The second question is about the spectral peak around 3 kHz found for some subjects. A similar spectral peak is also found in the voices of trained singers. Synthesis experiments made by Sundberg have shown that this peak is made up by three or four poles and not a single resonance. X-ray photos suggest that this spectral peak is caused by the shunting effects at the sinus piriformis which will add an extra pole around 3 kHz and a zero above that frequency. The zero may explain the fast fall in the spectrum above 3 kHz. Can this explanation also be used for the appearance of the peak in the voices of your subjects?

STEVENS

In reply to Jan Lindqvist's first question, I know of no experiments that investigate the use of the minimum around 1 kHz as a cue for speaker identification. An acoustic attribute such as the frequency of this minimum, which is presumably dependent on an anatomical feature not under the control of the speaker, should indeed demonstrate little intraspeaker variability. Mr. Lindqvist's second comment provides a more detailed possible explanation of the occurrence of a relatively fixed spectral peak around 3 kHz for some speakers. More acoustic data, coupled with measurements from X-ray pictures, are needed to give us a better understanding of the reason for this spectral peak.

SOVIJÄRVI (Helsinki)

As we know the intensity of the front and back vowel spectra have typical differences concerning the relative intensity of the fourth formant, too. Did you, Dr. Stevens, measure especially this cue? According to my experience, the *F*4 is very clearly dependent on the characteristics of the inter-individually varying voice quality. If the voice production is not air-tight, but e.g., soft, the *F*4 has a relative WEAK amplitude in all vowels and resonants of the speaker concerned. Have you had about the same experiences?

STEVENS

I agree that the amplitude of the fourth formant peak in the spectrum relative to, say, the amplitude of the first formant peak may provide an important cue for speaker identification, particularly for back vowels. This parameter could reflect the shape of the glottal spectrum at high frequencies and also the vocal-tract configuration in the region of the larynx, as discussed by Mr. Lindqvist in his remarks.

SMITH, S. (Hamburg)

Do you not think that for speaker identification a logarithmic display on our spectrograms would be able to show you more details in the analysis? Small deviations in the first formant region are often found in one speaker as against another one out of the same family.

STEVENS

A logarithmic scale would tend to give more space on the spectrographic display to the lower frequencies, and hence to accentuate those aspects that relate to the first few harmonics of voiced sounds and to the first formant. On the other hand, there may also be high-frequency information, not normally used to provide cues for linguistic units, that is relevant to speaker identification. Such information would be compressed in a logarithmic display, but not on a spectrum with a linear scale. It seems to me, therefore, that there are arguments for using both types of frequency scale.

HALLE (Cambridge, Mass.)

I do not think that this discussion should be allowed to come to a close without someone putting on record the fact that the question of speaker identification is not only an abstract scientific problem, but that it is also a topic in which various law-enforcement agencies especially in the U.S.A. have manifested a very lively interest. As readers of Solzhenitsyn's *First Circle* will no doubt recall, the U.S. law-enforcement agencies are following in this the precedent set by the Soviet MVD under Stalin, the only difference being that the Soviet MVD had its research work performed by scientists who had been imprisoned for political 'crimes', whereas our law-enforcement agencies are able to buy the needed research on the free market from scholars who presumably are interested in speaker identification as a purely abstract problem. This abstract interest in the problem completely dominated the discussion here: we talked of the effects of the zero due to the tracheal resonance, of the contributions of the fourth formant, of the influence of the *sinus piriformis*. No one here mentioned the fact that there is some danger that almost any result, no matter how tentatively and uncertainly put forward, may be put to practical use either by some law-enforcement agency or by one of the entrepreneurs who are supplying the police with all sorts of mechanical aids or are trying to persuade the latter that they have a burning need for the mechanical aid produced by the entrepreneur in question.

I do not wish to give the impression that I am opposed to helping the police in their often very difficult and dangerous task of apprehending thieves, robbers, and murderers. I cannot help but recall, however, that not long ago in Nazi-occupied Europe, the police apprehended hundreds and thousands of people who had not stolen, robbed or murdered, and similar things have been reliably reported to have happened in other places and times. Whether under these circumstances, one should work on providing the police with yet another weapon for their arsenal is a question to which some thought must be given. We live at a time when no one can fail to be impressed with the danger of scientific knowledge being misapplied. It behooves us, therefore, not to allow our natural curiosity about the effects of the sinus piriformis and about the role of the fourth formant to overlook the fact that by our work we might unwittingly be forging yet another instrument for the enslavement of men.

STEVENS

With regard to Professor Halle's comments, I can only share his concern that scientists should remain aware of the possible uses to which their findings are to be put. While we are engaged in research that can, hopefully, provide us with a better understanding of human communicative behavior, it is important that, when our findings have an influence on the non-scientific community, we communicate our results in a clear and impartial fashion that cannot be misinterpreted or misused.