Journal of Voice
Vol. 7, No. 2, pp. 97–117

Vocal Tract Acoustics

R. D. Kent

Waisman Center, University of Wisconsin-Madison, Madison, Wisconsin

Summary: This paper considers the topic of vocal tract acoustics from the three perspectives: (a) the acoustic theory of speech production; (b) contemporary laboratory methods for acoustic analysis, and (c) measurement of the acoustic signal of speech. Linear source-filter theory is the standard acoustic theory of speech production and is the foundation for remarkable advances in the analysis and synthesis of speech. Digital signal processing, the dominant laboratory method for speech analysis, enables the acquisition and recording of the acoustic speech signal but also implements quantitative algorithms largely based on linear source-filter theory. Measurements of the acoustic signal reflect the acoustic theory of speech production, laboratory methods for signal analysis, and principles of experimental phonetics. Basic issues in the three domains of theory, laboratory methods, and measurement are summarized as they pertain to the interests of the voice scientist, voice clinician, and voice teacher. **Key Words:** Vocal tract acoustics—Theory—Acoustic analysis—Acoustic measurement.

An understanding of vocal tract acoustics embraces three related areas that are reviewed in this paper: (a) the acoustic theory of speech production; (b) laboratory methods for acoustic analysis; and (c) measurements of the acoustic signal of speech. These three areas are strongly interrelated, as suggested by Fig. 1. Acoustic theory underlies the development of analytic tools and enables the interpretation of acoustic data. Laboratory instruments acquire and store the speech signal, and they often are designed to implement quantitative algorithms for acoustic analysis. An understanding of the modern acoustic analysis of speech requires, at minimum, an appreciation of linear source-filter theory (the standard acoustic theory of speech), digital signal processing (the heart of modern acoustic analysis), and the acoustic structure of the speech signal (which has been sufficiently well described to permit high-quality speech synthesis and reasonably good performance in machine speech recognition, at least for restricted conditions).

This paper will review basic issues in these areas and highlight modern developments, particularly as they apply to concerns of the voice scientist, voice clinician, and voice teacher. Swift and profound advances in acoustic analysis of speech have greatly increased the power and availability of this tool for a broad range of users. It is not possible to consider these developments in anything beyond broad strokes in this paper, but references are given to more detailed discussions of the major topics. In this sense, the present paper is a guide to the contemporary issues and literature on the acoustic analysis of speech, with emphasis on vocal tract acoustics. The paper assumes only a general background in acoustics and speech production.

ACOUSTIC THEORY OF SPEECH PRODUCTION

The contemporary understanding of vocal tract acoustics is based almost entirely on a linear, time-invariant source-filter model. The standard reference is Gunnar Fant's (1) "Acoustic Theory of

Accepted November 15, 1991.
Address correspondence and reprint requests to Dr. R. D. Kent at Waisman Center, University of Wisconsin-Madison, 1500 Highland Ave., Madison, Wisconsin 53705-2280.
This paper was presented at the Voice Foundation's 20th Anniversary Symposium: Care of the Professional Voice, July 16, 1991.

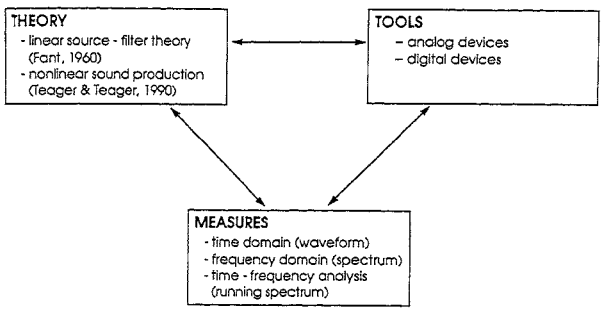

FIG. 1. Three interrelated areas pertaining to the understanding and application of vocal tract acoustics.

Speech Production." This theory has been remarkably productive and has received few serious challenges (but see Teager and Teager (2) for an alternative view). An understanding of the source-filter theory is an excellent foundation for interpreting phenomena of voice and speech. The assumptions of linearity and time invariance make the acoustic analysis of speech tractable. Simply put, linearity states that the system of interest obeys the superposition principle, meaning that the response of the system to a sum of simple inputs is the response to the sum of those inputs. Time invariance means that the response of the system to a time-delayed or time-advanced input is similarly time-delayed or time-advanced. These assumptions make possible the application of a set of powerful analytic techniques to the examination of the acoustic signal of speech. As will be discussed, these techniques are available on a number of speech analysis systems that run on microcomputers.

The source-filter concept proposes that acoustic energy generated by a sound source is passed through a frequency-dependent transmission system. The task of speech analysis therefore is largely one of identifying a sound source and describing a corresponding filter function. There are three major sources to be considered: (a) laryngeal voicing source, typified in the phonation of vowels; (b) turbulence noise source as in the case of the fricative consonants; and (c) transient source, which applies to the release burst of stop consonants. With appropriate modifications, these three sources account for the various classes of sounds that make up the phonetic system of English (and many other languages as well).

Source-filter theory for vowels

The source-filter theory for vowel production is illustrated in Fig. 2 and summarized for a frequency domain analysis by the following formula:

$$P(f) = U(f)\ T(f)\ R(f), \qquad \text{Eq. (1)}$$

where (f) indicates a function of frequency, f,
P(f) is the radiated sound pressure spectrum,
U(f) is the glottal volume spectrum,
T(f) is the vocal tract transfer function, and
R(f) is the radiation characteristic.

Basically, equation 1 states that the sound pressure, as might be measured by a microphone placed near a speaker's mouth, is the product of the glottal volume velocity (the source energy), the vocal transfer function (part of the filter function) and the radiation characteristic (another part of the filter function, relating the volume velocity spectrum of the source to the sound pressure spectrum of the radiated signal). The multiplication of the terms in the frequency domain is equivalent to the mathematical operation of convolution in the time domain. This discussion will emphasize the frequency domain analysis, as expressed in Eq. (1). However, each term in Eq. (1) can be related to a corresponding time function, e.g., P(f) in the frequency domain and p(t) in the time domain.

The source of acoustic energy for vowels is typically the voicing signal generated by the vibrating vocal folds. The glottal spectrum, U(f), and the corresponding glottal volume-velocity waveform, u(t), are shown in their well-known idealized forms in Fig. 3. The idealized (and simplified) laryngeal waveform is a series of triangular pulses spaced at the fundamental period, To, the reciprocal of the fundamental frequency, fo. The laryngeal spectrum, or Fourier transform of the waveform, is a

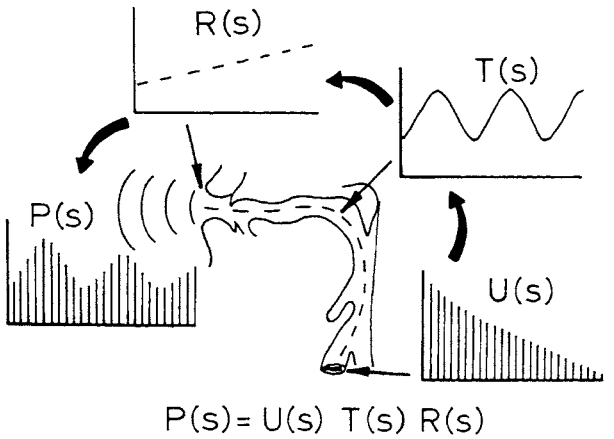

FIG. 2. Diagram of the vocal tract showing the affiliation of vocal tract regions with the major terms of the source-filter theory.

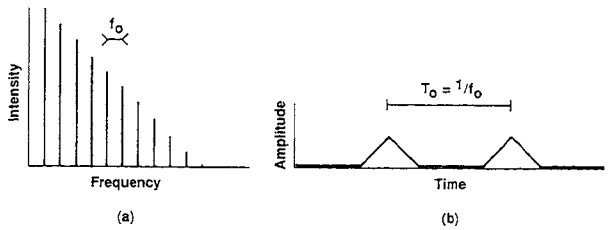

FIG. 3. a and b: Idealized form of the glottal spectrum, U(f), and the associated waveform, u(t).

harmonic spectrum in which the components diminish in amplitude at the rate of 12 dB/octave. Mathematically speaking, the amplitude reduction of the successive harmonics is necessary for convergence of the Fourier transform. Practically speaking, the spectrum shows that voicing energy is dominated by low-frequency harmonic components. The 12 dB/octave figure is a mathematical ideal. Actual relationships among harmonic amplitudes vary with laryngeal configurations and speaker variables including pathology. Variations in fundamental frequency and intensity introduce systematic changes in the laryngeal waveform and spectrum.

The acoustic pulses generated by the vibrating folds propagate through the vocal tract, where filtering occurs. Although introductory accounts usually assume that source and filter are independent, it is now abundantly clear that this assumption is not correct, as source and filter do interact. The nature of this interaction is a topic of considerable current interest (3) and it carries important implications for speech and singing.

The filter function, or transmission function, of the vocal tract is defined primarily by the resonances of the vocal tract. These resonances are called formants, and each formant is specified by a center frequency (formant frequency, abbreviated Fn, where n is the formant number) and a bandwidth. Formant amplitude might be included as well, but in general, relative formant amplitude can be derived from formant frequency and bandwidth information. Theoretically, there are an infinite number of formants, but no more than 3–5 usually are considered in speech analysis. The formants are resonant properties of the vocal tract. They can be determined mathematically from a precise knowledge of the vocal tract shape or they can be estimated from measurements of the acoustic signal.

Mathematical prediction of formants is possible given that the formant frequencies depend on the length of the vocal tract and the cross-sectional shape of the vocal tract as a function of its length. These two variables, length and cross-sectional shape as a function of length, are conveniently expressed in graphic form as the vocal tract area function (Fig. 4). Vocal tract length determines the average spacing of formant frequencies. This follows from a simple acoustic model, a tube closed at one end and open at the other, as shown in Fig. 5. The closed end refers to the vocal folds and the open end to the mouth opening. If we assumed the simplest case in Fig. 4, then the cross-sectional area is uniform over the length of the tract. In this case, the formant frequencies are determined only by the length of the tube according to the odd-quarter wavelength relationship:

$$Fn = (2n - 1)\, c/41 \qquad \text{Eq. (2)}$$

where Fn is a particular formant frequency,

(2n − 1) gives the odd intergers,

c is the velocity of sound, and

l is the length of the vocal tract.

As l becomes smaller, the value of any particular Fn will increase. Conversely, as l becomes larger, the value of any particular Fn will decrease. Therefore, formant frequencies vary with the length of the vocal tract and therefore with speaker characteristics such as age and sex. Furthermore, within persons of a given sex, formant frequencies vary with vocal tract dimensions. Tenors and basses differ in formant frequency patterns in much the way that females differ from males (4). That is basses have lower formant frequency values than tenors.

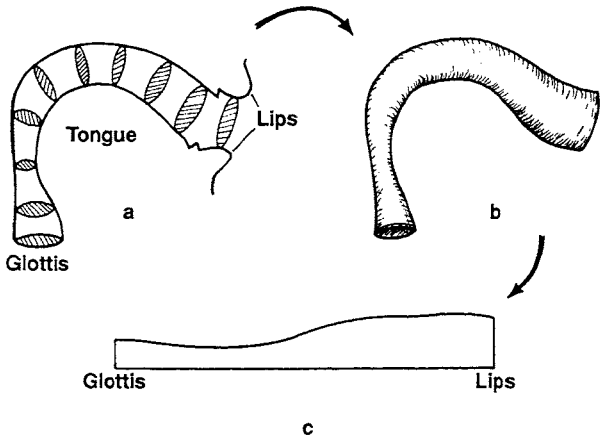

FIG. 4. Vocal tract area function shown as (a) curved vocal tract with selected points of cross-dimension measurement, (b) derived area function for a curved tube, and (c) area function for an equivalent straight tube.

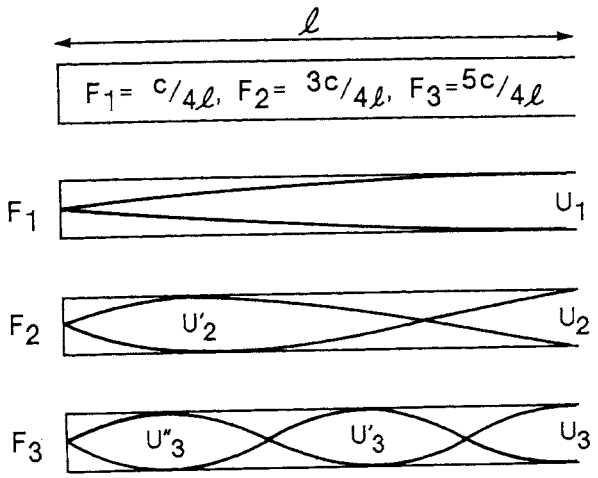

FIG. 5. Straight tube closed at one end (glottis) and open at the other (lips), showing stationary distribution of volume velocity for the first three formants, F1, F2, and F3. The resonances of the tube are given by the odd-quarter wavelength relationship (a tube of this configuration will resonate with maximal intensity to a sinusoid whose wavelength is four times the tube length).

The length l of the vocal tract also varies within a speaker as the result of lip protrusion and larynx lowering, both of which extend the vocal tract length. Lip protrusion is an articulatory feature that accompanies certain vowels, especially most back vowels in English. Larynx lowering or raising is not as commonly regarded as a phonetic characteristic, but it may in fact be used by many speakers and singers. For example, some baritones who have the capability to change their voice category to tenor may accomplish this change by laryngeal elevation (4).

Resonance phenomena for the tube in Fig. 5 can be discussed in terms of the stationary distribution of volume velocity, or its inverse, pressure. Distinct regions of volume-velocity maxima and minima arise in the tube, as illustrated in Fig. 5. These volume-velocity distributions result from interactions of particle vibration within the tube (1).

The simple tube model in Fig. 5 must be modified to apply to different vowels. In particular, the model must allow the cross-sectional area to vary over the length of the vocal tract. The area functions are derived by measuring the vocal tract cross dimensions at selected points and then plotting these values as a function of length. In so doing, the curved vocal tract is straightened out. This change in geometry (straightening) has little effect on the formant frequencies (5).

The stationary distribution of volume velocity is basic to the perturbation theory of articulatory-acoustic relationship. This theory proposes that a perturbation of the vocal tract configuration, i.e., a local narrowing in the area function, causes predictable changes in formant frequencies, depending on the proximity of the narrowing to a volume velocity maximum or minimum. The general rules are quite simple: (a) a perturbation near a volume velocity maximum for a given formant causes the frequency of that formant to decrease; and (b) a perturbation near a volume velocity minimum for a given formant causes the frequency of that formant to increase.

If formant frequencies are generally predictable from perturbation theory, then can formant amplitudes also be predicted? Fant (1) showed that they can, and some general rules for determining formant amplitudes are as follows: (a) Increasing (decreasing) the frequency of F1 causes the amplitudes of higher formants to increase (decrease). (b) When two formants move closer together (farther apart), their amplitudes increase (decrease). These relations follow from the interaction of formants and can be conceptualized as the graphic addition (in a dB scale) of separate resonance curves to form the overall transfer function of the vocal tract.

This point is relevant to the so-called singer's formant, which Sundberg (4) describes as a peak in the spectrum between 2 and 4 kHz, depending on voice type (low-frequency peak for basses, high-frequency peak for tenors). The peak is the consequence of formant tuning such that higher formants assume frequencies close to that of F3, in a kind of formant clustering. As noted in the general rules above, a close tuning of formants tends to increase their amplitudes. Therefore, a clustering of higher formants in the vicinity of F3 will yield a spectral peak. This adjustment is an example of tuning, in which a singer changes the vocal tract configuration and/or vocal fold function to achieve a particular acoustic result.

Linear source-filter theory has been a highly productive theory. It has had substantial impact not only as a conceptual framework but also in practical matters such as acoustic phonetic description and the development of speech synthesizers. To be sure, linear theory is only an approximation. It applies to frequencies for which longitudinal propagation can be assumed. Inaccuracies occur for frequencies at which cross-mode vibrations occur in the vocal tract. Fujimura (6) estimates that this fre-

quency can be as low as 1,700 Hz for a cross dimension greater than 5 cm. It usually is assumed that cross-mode vibrations are minimal below frequencies of about 5 kHz.

Teager and Teager (2) presented evidence for the existence of important nonlinearities, even for vowels. They describe nonlinear processes related to the nonlinear interaction of sheet jet flows and generated flow vortices. Teager and Teager make the provocative statement that "the operation of the vocal tract is neither linear nor passive, nor even acoustic" (2). This work is very interesting, but Fujimura (6) argues that it doesn't appreciably detract from the remarkable success of the standard linear theory. The standard theory has been confirmed in important respects experimentally (7), has been the basis for the development of successful formant synthesizers (8,9), and underlies contemporary methods of acoustic analysis, such as linear predictive coding (LPC) (10) and cepstral analysis (11). Finally, as Fujimura (6) points out, turbulence effects involve DC airflow, which is excluded from acoustic measurements.

But it is also important to note Fujimura's (6) remark that the general success of the standard theory does not mean that there is no room for improvement. For example, because source-filter independence often cannot be assumed, interactions of source and filter are being examined for speech, singing and whistling. (For a concise discussion of glottal flow models and source-filter interaction, see Fant (3).) Another direction for revision of the theory pertains to dynamic considerations, e.g. consideration of inertial effects in articulatory movement. Finally, the application of the standard acoustic theory makes a number of assumptions, the validity of which should be evaluated for individual applications. Assumptions that are safely made in one situation may not apply equally well to other circumstances.

Vowel formant patterns

A primary conclusion from research based on the standard acoustic theory is that a particular vowel is associated with a characteristic formant pattern, as illustrated in Fig. 6 for the first two formant frequencies, F1 and F2. This illustration shows stylized spectrograms for vowels classified with respect to their articulation as front, central, and back (anteroposterior placement of the tongue) and along a low-high continuum (superoinferior placement of the tongue). Generally, the F1 frequency varies strongly with tongue height (or jaw opening), and the F2 frequency varies more with the position of the tongue in the anteroposterior dimension.

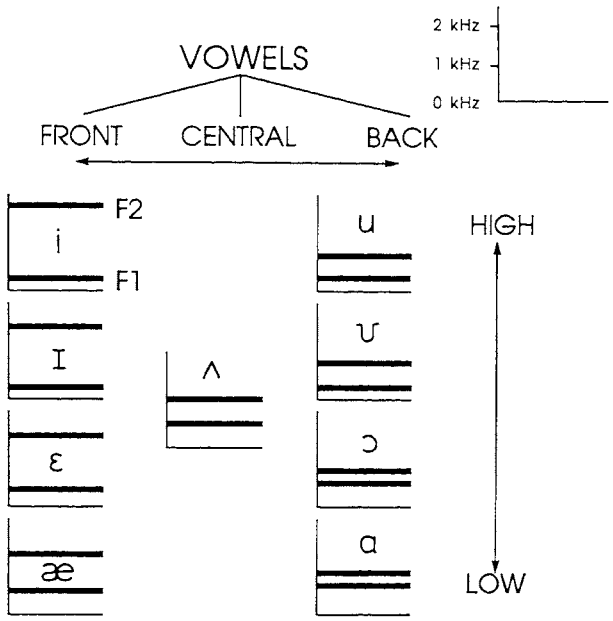

FIG. 6. Acoustic-articulatory relations for vowels. Front vowels are associated with a fairly wide F2–F1 separation, back vowels with a narrow F2–F1 separation. Therefore, F2–F2 separation correlates with advancement or retraction of the tongue. High vowels are associated with a low F1, low vowels with a high F1. Therefore, F1 frequency correlates with tongue height (or jaw opening). The effect of lip rounding, not shown, is to lower all formant frequencies. In English, only the back vowels and r-colored vowels are rounded.

Knowledge of these articulatory-acoustic relationships can be helpful for both theoretical and practical considerations. As a practical example, many devices or software programs that extract the vocal fundamental frequency from the speech signal employ a severe high-pass filtering to eliminate the influence of formants. The resulting signal is a simplified waveform that roughly matches the laryngeal waveform (Fig. 3). A limitation on this technique is that the filter should be set to exclude the lowest-frequency formant, F1, but not the fundamental frequency. The ideal vowel, then, is one in which F1 is widely separated from the fundamental frequency. As Fig. 6 shows, the vowel with the highest F1 frequency is vowel /ɑ/, a low-back vowel. The same vowel is often used in visual examination of the throat, since the low-back position of the tongue and the large opening of the jaw facilitates direct

viewing with the unaided eye or with a laryngeal mirror or other optical device.

The idea that a particular vowel is associated with a distinctive formant pattern gave rise to a Target theory of vowel recognition. This theory proposes that a vowel's formant pattern is sufficient for identification of the vowel. Presumably, then, a vowel can be adequately specified with a static formant pattern, such as the frequencies of the first three formants. This theory has been questioned because of evidence that dynamic (temporal) factors can play a large role in vowel identification (12). It is likely that both static features (relatively long-term features such as stable formant pattern) and dynamic features (such as variations in formant pattern around the vowel steady-state) contribute to the identification of vowels. Vowels are rather elastic in the sense that a vowel often can be produced with different durations. For example, in singing, vowels may be prolonged along with their accompanying notes. In speech, vowel durations are markedly affected by factors such as stress and speaking rate.

The formant patterns shown in Fig. 6 are typical of the speech of adult males. Figure 7 shows vowel ellipses drawn in the F1–F2 vowel diagram to enclose the F1–F2 values for a given vowel produced by men, women and children. This figure is based on data presented in the classic report of Peterson and Barney (13), which continues to be cited as a primary source of data on vowel formant frequencies. Values from the Peterson and Barney report are given in Table 1. Because vowel formant frequencies vary with the length of the speaker's vocal tract, normalization of formant frequencies across speakers has been a major problem in acoustic research (14–18).

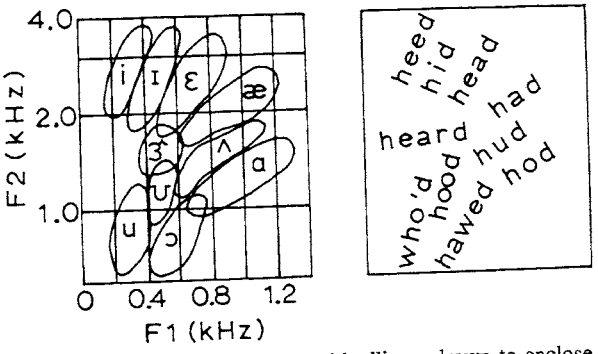

FIG. 7. **Left:** F1–F2 vowel chart with ellipses drawn to enclose the data for a large group of men, women and children. Values for men are at the end of the ellipses closest to the origin, values for women are close to the middle of the ellipses, and values for children are at the end of the ellipses farther from the origin. **Right:** The accompanying graph shows the approximate location of keywords for each vowel phonetic symbol shown in the ellipses in (a).

F1–F2 correlates of vowel articulation are summarized in a following section on acoustic measures of speech. Discussions of speech acoustics in basic textbooks usually emphasize the lowest two or three formants (F1, F2, and F3). Because these formants are most important for the identification of vowels, they are often discussed at the exclusion of the higher formants. But the higher formants are not negligible for all purposes. In the case of speech, it was recognized in attempts to synthesize speech with machines that inclusion of the higher formants adds naturalness to the speech, even if these formants assume invariant values across different vowels. In singing, the higher formants also add distinctiveness. Sundberg (4) drew attention to F4 in regard to voice timbre, the personal component of voice sound.

Relating vocal tract shape for vowels to acoustic output

A fundamental problem in speech acoustics is to derive the acoustic output from the vocal tract shape, or conversely, to derive the vocal tract shape from the output signal. The problem is complicated by the fact that an infinite number of vocal tract shapes theoretically could be associated with a particular output spectrum. Therefore, efforts have been directed at constraining the possible shapes that can be assumed by the human vocal tract. A related interest is to derive a description of the vocal tract configuration that is simpler than a detailed area function. The area function is potentially complex insofar as the cross-sectional area must be specified for the length of the tube (about 17 cm in men). Even if one assumed that the tube could be approximated by a series of sections of equal length (say, one cm), 17 cross-sectional areas would be required. Is there a simpler way, preferably one with direct articulatory-acoustic relationships?

One general approach to simplified description is constriction parameterization. The idea is to give a general description of the vowel articulation. Stevens and House (19) (see also Fant (1)) proposed such a solution in their three-parameter model of the vocal tract shape for vowels. The three parameters were: (a) location of the constriction; (b) size of the constriction; and (c) the ratio of mouth open-

TABLE 1. *Formant frequencies (in Hz) of the first three formants (F1, F2, F3) of ten vowels produced by 76 speakers including men, women and children (values drawn from Peterson and Barney, 1952)*

	Men			Women			Children		
Vowel	F1	F2	F3	F1	F2	F3	F1	F2	F3
[i]	270	2,300	3,000	300	2,800	3,300	370	3,200	3,700
[I]	400	2,000	2,550	430	2,500	3,100	530	2,750	3,600
[ɛ]	530	1,850	2,500	600	2,350	2,000	700	2,600	3,550
[ae]	660	1,700	2,400	860	2,050	2,850	1,000	2,300	3,300
[*a*]	730	1,100	2,450	850	1,200	2,800	1,030	1,350	3,200
[ɔ]	570	850	2,400	590	900	2,700	680	1,050	3,200
[U]	440	1,000	2,250	470	1,150	2,700	560	1,400	3,300
[u]	300	850	2,250	370	950	2,650	430	1,150	3,250
[ʌ]	640	1,200	2,400	760	1,400	2,800	850	1,600	3,350
[ɜ˞]	490	1,350	1,700	500	1,650	1,950	560	1,650	2,150
Mean	500	1,420	2,400	575	1,700	2,800	670	1,900	3,250
F2/F1	2.84			2.96			2.84		
F3/F2		1.69			1.65			1.71	

Values for F2 and F3 have been rounded to nearest 50 Hz.
The vowel means may be taken to define the approximate formant frequencies of a neutral vowel for each group.
Mean F2/F1 and F3/F2 ratios are shown at the bottom of the table.

ing to length. Stevens and House derived nomograms relating F1, F2, and F3 frequencies to these articulatory parameters. This work demonstrates that the primary acoustic features of vowels are described quite well by just three articulatory parameters: one specifying the location of the major constriction; one to indicate the size of this constriction; and another to gauge the degree of mouth opening. From the point of view of a speaker or singer, then, the object is to adjust these three parameters in accord with the intended phonetic quality. Nomograms relating the frequencies of the first five formants to the three control parameters of a four-cavity model of vowel production are available to Fant (1). Badin, Perrier, Boe and Abry (20) examined similar nomograms for "focal points"—regions where formant convergences occur and where formant-cavity affiliations are exchanged.

A particular advantage of these nomograms, and the principles on which they are based, is that they can be used to understand some articulatory compensations and certain adjustments of the vocal tract in speech and singing. Compensations are used by disordered speakers and by speakers who want to produce speech with unusual production patterns (such as ventriloquists, who try to avoid visible mouth movements). Some speakers with physical or neurological injury to the vocal tract can learn to produce adequate speech by using unusual articulatory patterns. Singers may rely on these principles to achieve adjustments of formant structure, as in efforts to tune F1 with the vocal fundamental frequency or in production of the singer's formant (4).

The statistical approach of factor analysis also has been taken to parameterize articulator-acoustic relationships (21–24). The object is to determine the smallest number of factors that satisfactorily account for variations in articulatory configuration. The factor analytic studies typically indicate that vowel articulation can be parameterized as two tongue factors, a lip factor and a jaw factor. This result represents a powerful simplification over a detailed articulatory description that specifies the shape of various portions of the vocal tract. It shows that vowel articulation may be described with only four articulatory parameters.

Another approach is to discover the acoustic consequences of the movements of individual articulators (25–27). A particular advantage of this work is that it addresses the way in which adjustments of a given articulator, such as tongue, lips, or jaw can effect the acoustic signal. For example, when tongue position is held constant, jaw motion affects primarily the value of F1 frequency. When jaw position is held constant and the tongue moves in the anteroposterior dimension, the primary changes are in the F2 frequency. Such articulatory-acoustic relationships are important in understanding how changes in articulatory position will affect the acoustic signal in speech or singing. Sundberg (4) gives several good examples in singing. One of these is the adjustment that female opera singers make in the frequency of F1. Sundberg observed

that the singers tend to have a greater jaw opening for higher notes than for lower notes and explained this result as follows: As fundamental frequency increases, it may exceed the frequency of the first formant, thereby reducing the sound level of the vowel. The singer compensates for this effect by lowering the jaw, which increases the frequency of F1. In this way, the singer tunes the F1 frequency to the fundamental frequency of the voice. Singers may use several such techniques. Acoustic theory enables these techniques to be understood in a coherent and parsimonious way.

Stevens (28) described what he called the "quantal nature of articulation," or the idea that nonlinearities exist in the relationship between vocal tract configuration and the acoustic signal. Stevens assumed that changes in articulatory parameters are not necessarily accompanied by commensurate changes in the acoustic signal. To the contrary, the quantal nature of speech is based on the assumption that articulatory adjustments and their acoustic effects have pronounced nonlinearities. These nonlinearities define critical regions of the vocal tract in which small articulatory adjustments can produce relatively large acoustic consequences. These regions presumably would require precise control of articulation to achieve a desired acoustic result. Therefore, languages would tend to avoid such critical regions as places of articulation for speech sounds. These regions also would be those that a speaker or singer protects from undue articulatory variation.

Support for Stevens' ideas can be found in the work of Wood (29) and Perkell and Nelson (30), who showed that tongue articulations for vowels vary least in the dimensions that are most critical for acoustic output. In particular, Wood showed that there are four constriction locations that can be related to a definable class of vowel qualities. These regions are: (a) along the hard palate; (b) along the soft palate, (c) in the upper pharynx, and (d) in the lower pharynx. Thus, in the supralaryngeal system, there may be acoustically determined constraints operating to control the amount of variation in articulatory positions and perhaps to guide the selection of elements in a phonetic system. The regions just described would seem to allow variation in mouth opening, and this point is relevant to the understanding of vocal-tract adjustments in speech and singing. For example, singers may be encouraged to hold the jaw in a relatively open position. If a certain degree of mouth opening were critical to vowels, then such instruction would result in unintelligible speech. However, if mouth opening is not as critical as other features of the vocal tract configuration, intelligible sound may be produced even with large variations in jaw opening.

Proposals similar to the quantal theory are incorporated in Carre and Mrayati's (31) "Distinctive Regions and Modes" theory and in Badin et al. (20) hypothesis of focal points in the articulatory-acoustic conversion. The common feature in these proposals is the idea that the acoustic signal is highly sensitive to adjustments in certain regions of the vocal tract. Presumably, a speaker or singer has implicit knowledge of these nonlinearities.

This discussion has emphasized articulatory-acoustic relations for the first two or three formants. But the higher formants can be similarly treated. For example, F4 is related most strongly to the length of the vocal tract and the dimensions of the vocal tract in the vicinity of the larynx tube. For additional discussion of higher formants in singing, see Sundberg (4).

Source-filter theory for consonants

Consonants are classified according to their production characteristics, typical place of articulation, manner of production, and voicing. Consonants have lower sound levels than vowels but they contribute significantly to intelligibility. Acoustic theory for consonants can be summarized in terms of manner of articulation categories.

Nasal consonants are made with a complete constriction in the oral cavity and an open nasal tract permitting nasal radiation of sound energy. A simple model is shown in Fig. 8. The acoustics of nasalization are actually rather complicated (Fujimura (32)), but three principle effects are observed: (a) the bifurcation of the tract introduces antiformants (or zeros); these have essentially the opposite effect of formants, causing acoustic energy in the region of the antiformant to be short circuited within the vocal tract; (b) a strong low-frequency nasal formant arises in strong association with the tube extending from larynx to nares; for men's speech, the nasal formant typically occurs at a frequency of about 300 Hz; and (c) the fleshy, convoluted lining of the nasal passages absorb a considerable amount of acoustic energy; this large damping is reflected in a broadening of formant bandwidths and a reduced overall energy.

The perception of nasal consonants involves an integration of murmur (the acoustic interval associ-

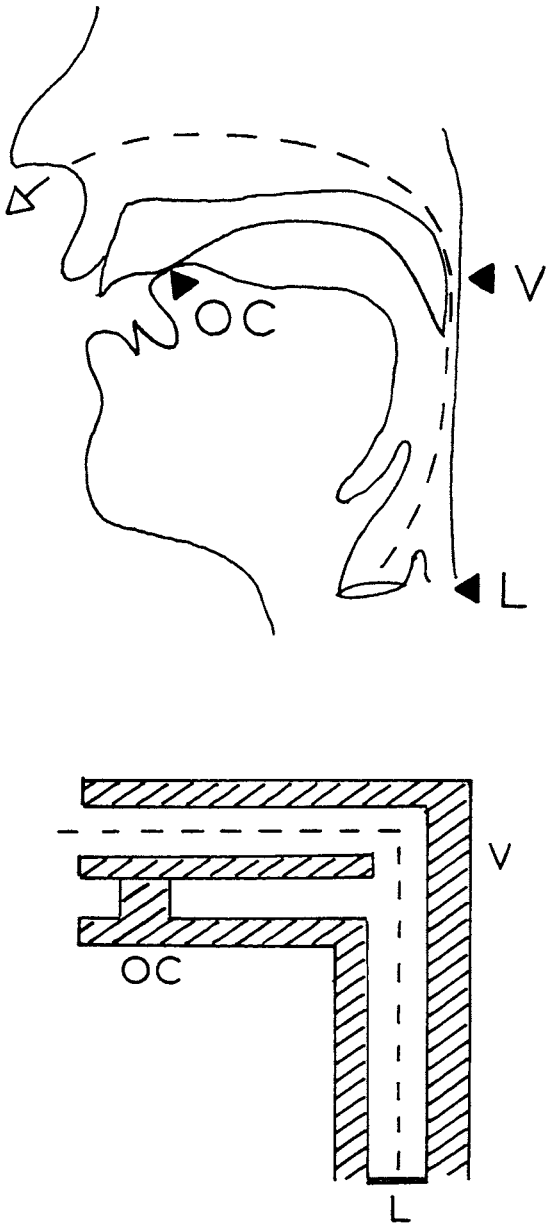

FIG. 8. Model of vocal tract for a small consonant. The oral cavity is constricted (OC, oral constriction) so that sound energy passes through the nasal cavities. V, velum; L, larynx.

ated with nasal radiation of sound energy) and transition cues (33,34). Nasal consonants (and vowels that become nasalized owing to the influence of neighboring nasal consonants) have complicated acoustic properties, and it should not be expected that one invariant change in the acoustic signal will identify nasalization. An important point to remember is that nasalized sounds (both consonants and vowels) are weaker than oral vowels, which are usually the most intense component in a string of sounds.

Fricative consonants have as their essential feature the generation of turbulence noise. Noise is produced at a region of vocal tract constriction. As shown in Fig. 9, the constriction acts like a nozzle so that air exiting from it forms a jet. As the jet mixes with surrounding air, eddies form in the flow in the vicinity of the contraction and expansion of the constriction. The eddies are rotating volume elements of air, that is, irregular, high-frequency fluctuations in velocity and pressure at a point in space. For a constriction of given dimensions, turbulence noise is generated at a critical flow velocity given by the well-known Reynold's number (for general discussion of these principles, see Kent and Read (35) and Shadle (36)). The Reynold's number, Re, is defined as:

$$Re = vh/\nu \qquad \text{Eq. (3)}$$

where v = flow velocity (cm^3/s)

ν = kinematic coefficient of viscosity (about 0.15 cm^2/s for air), and

h = characteristic dimension (for flow through an orifice, h is on the order of the diameter of the orifice).

Turbulence can be visualized in a slow motion film of the events that occur as a colored fluid is injected into another fluid. As Re increases gradually, one could observe an initial region of laminar flow (smooth layered air movement), then an unstable interval, and finally a condition of full turbulence marked by the formation of eddies.

Because volume flow, U(cm^3/s), is determined as

$$U = vA \text{ (A is cross-sectional area)}, \qquad \text{Eq. (4)}$$

the Reynold's number also can be calculated as

$$Re = Uh/A\nu. \qquad \text{Eq. (5)}$$

MODEL OF TURBULENCE NOISE PRODUCTION FOR FRICATIVES

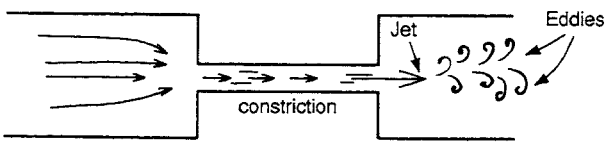

FIG. 9. Model of vocal tract for turbulence noise generation. Air forced through the constriction forms a jet at its outlet. Eddies (rotating volume elements of air) are associated with turbulence.

The volume flow U depends on the constriction size and the pressure differential across the constriction, Ps:

$$U = kA\sqrt{Ps} \text{ (where k = constant).} \quad \text{Eq. (6)}$$

Then

$$\begin{aligned} Re &= Uh / A\nu \\ &= kA\sqrt{Ps}\, h / A\nu \\ &= kh\sqrt{Ps} / k\nu \end{aligned} \quad \text{Eq. (7)}$$

Thus, the value of the Reynold's number, which corresponds to turbulent flow if the value is sufficiently high, depends on the size of the constriction and the pressure driving the air through the constriction.

Turbulence is the energy source for fricatives, the frication portion of affricates, the burst of stops and the breathiness or aspiration produced in a narrow glottis. What the listener hears as noise is the energy produced by random pressure fluctuations of the turbulent field. Volume velocities for fricative consonants lie in the range of 100–1,000 cm/s. The critical Reynold's number for speech noise is Re > 1,800.

Shadle's (36) modeling studies showed that there are at least two major ways in which fricative noise is generated. The first, involving an obstacle source, generates sound in the region of a rigid body approximately normal to the flow. An example is the palatal fricative /š/, for which the lower teeth form the obstacle (which is like a spoiler in a duct). According to Shadle, an obstacle source is associated with a maximum source amplitude for a given flow velocity, by a relatively flat spectrum that falls off with increased frequency, and by a maximum rate of change of sound pressure with volume velocity. It is well known that young children who lose their central incisors until their permanent replacements appear have altered fricative sounds. A major reason for this affect is the loss of the spoiler formed by the central incisors.

The second noise source involves a wall source, which occurs when sound is generated primarily along a relatively rigid wall that runs roughly parallel to the flow. The fricatives /ç/ and /x/ are examples of this kind of source, which is associated with a high (but less than maximum) source amplitude for a given flow velocity, by a spectrum that possesses a broad peak, and by a high (but not maximum) rate of change of sound pressure with volume velocity. Shadle proposed that the wall source is a distributed source, whereas the obstacle source can be modeled as a series pressure source located at the obstacle.

Most fricatives can be modeled as a two-cavity tube formed as the constriction is made at a point corresponding to the place of articulation (Fig. 10). Generally, the spectral shaping of the fricative is determined by the resonances of the front cavity. The lowest resonance is given by the odd-quarter wavelength relationship already discussed for vowels. For example, if the front cavity has a length of 2 cm, the calculated resonance would be about 4,000 Hz. However, the back cavity resonances also become relevant under certain conditions, as when the back cavity is tapered so as to produce a gradually narrowing constriction. Stevens (28) discussed fricative production in terms of the quantal nature of the articulatory to acoustic relationships. A number of reports have been published on the spectral properties of fricatives produced by children (37–41).

Turbulence noise is important not only for the understanding of fricatives and other noisy consonants, but also for an understanding of whistling and breathy voice quality. Breathiness arises as turbulence noise is generated in the larynx and possibly the lower pharynx. When this turbulence noise is mixed with the periodic vibration of the vocal folds, the result is breathy phonation.

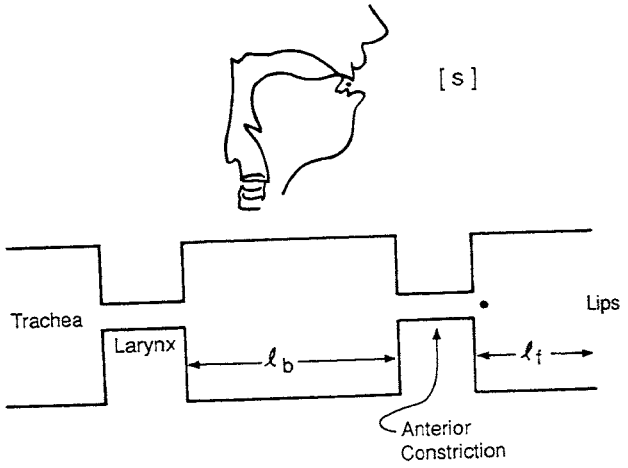

FIG. 10. Model of vocal tract for fricative production, with trachea, constriction at larynx, back cavity, articulatory constriction, and front cavity. The dimensions of back cavity length and front cavity length can be used to estimate resonance properties. Example shown is (s).

Stop consonants (/b d g p t k/) must be described with respect to potential cues, given that the acoustic cues for a stop vary with context and syllable position. Generally, the cues are some combination of the following: (a) stop gap: an acoustic interval corresponding to vocal tract closure. This interval is truly silent for voiceless stops but often contains voicing energy (evident in a spectrogram as a low-frequency voice bar); (b) release burst: this transient energy, usually of no more than 20–30 ms duration, is generated as the articulatory constriction is released; and (c) formant transitions: as the occluding articulator moves away from the consonant toward another vocal tract configuration (especially that of a vowel), the formants shift in frequency over an interval of about 50 ms.

Despite its brevity, the stop burst seems to be highly informative regarding place of articulation. Numerous papers have been published on the relative information carried by the stop burst and formant transitions (42–49). Recent studies indicate that the burst carries substantial information regarding place of articulation. It also appears that the dynamic properties (time-varying spectral features) of the burst are highly important aspects. However, if these temporal features are neglected in the interest of formulating a general rule for burst spectra, the following principles are useful: (a) the bilabials [p b] have a diffuse-falling spectrum, meaning that the noise energy is widely distributed over the frequency range and that the overall spectrum has most of its energy in the low-frequencies (therefore, the spectrum "falls" with increasing frequency); (b) The alveolars [t d] have a diffuse-rising spectrum, meaning that the noise energy is widely distributed with most of the energy in the high frequencies (therefore, the spectrum "rises" with increasing frequency); and (c) the velars [k g] have a compact spectrum (dominant midfrequency peaks). The word *compact* implies a concentration of acoustic energy in the midfrequency region.

Formant transitions are less easily characterized, as they depend on both the formant loci for the consonant and the formant pattern for the vowel. Klatt (50) is a good source on the temporal patterns for burst, frication and voice onset time in syllable-initial stops.

Affricate consonants [tʃ] and [dʒ] are combinations of stop and fricative articulations. Like stops, the affricates have an interval of vocal tract closure. Like fricatives, these sounds have a frication interval. Affricates stand in intermediate position between stops and fricatives in that the noise portion of affricates is longer than that for stops but typically shorter than that for fricatives. It has been proposed that a distinguishing feature between affricates and fricatives is that the former have a shorter rise time of the amplitude envelope (51).

Liquid consonants are the lateral [l] and the rhotic [r]. These sounds are often troublesome to children and tend to be mastered late in phonetic development. The [l] resembles a nasal consonant in having both formants (poles) and antiformants (zeros) in its transfer function. The zeros arise because of the bifurcation of the vocal tract created by the midline apical constriction for [l]. The [l] has a formant structure characterized by a low-frequency F1, midfrequency F2, and a high-frequency F3. Mean formant frequencies for [l] reported by Nolan (52) were: F1, 360 Hz; F2, 1,350 Hz; and F3, 3,050 Hz.

The [r] has a well defined formant pattern, the distinguishing feature of which is a small F3–F2 difference. For example, Nolan (52) reported the following mean formant frequencies for [r]: F1, 320 Hz; F2, 1,090 Hz; and F3, 1,670 Hz. The F1 and F2 frequencies are similar to those for [l], but the F3–F2 difference is about 600 Hz for [r] compared to about 1,500 Hz for [l].

The acoustic properties of the liquids also have been described in several other papers (53–56).

Glide consonants (semivowels) are the [j] and [w]. These have well-defined formant patterns in which formant frequencies change gradually over an interval of about 60–100 ms. The formant pattern for [j] is similar to that for the high-front vowel [i] (which [j] resembles in articulatory configuration). The formant pattern for [w] is similar to that for the high-back vowel [u] (which [w] resembles in articulatory configuration). Discussions of glides are available in Liberman et al. (55) and O'Connor et al. (56).

In his discussion of the relations between the vocal tract area functions and the acoustic signal of speech, Fant (57) summarizes the research agenda of vocal tract modeling: ". . . to improve techniques for inferring vocal tract characteristics from speech wave data we need a better insight into vocal tract anatomy, area function constraints, and a continued experience of confronting models with reality—a balanced mixture of academic sophistication and pragmatic modeling" (57).

LABORATORY INSTRUMENTS FOR SPEECH ANALYSIS

For many years, the central piece of equipment for the acoustic analysis of speech was the sound spectrograph, which produced a three-dimensional (intensity × frequency × time) analysis in the form of the spectrogram. The spectrograph was the origin of much of the data that made acoustic phonetics a laboratory science. A useful reprint collection of 33 papers on speech spectrography was recently published (58). It includes papers in the categories of: Basics and Beginnings, Spectrographic Characteristics of Normal Speech, Speech Sound Development, and Spectrography in Evaluation and Therapy.

The original spectrographs operated on the analog voltage signal of speech to produce a running short-spectrum, which was printed in hard copy form by a controlled burning of facsimile paper. The blackness of the burning produced the gray scale by which intensity was represented.

The modern acoustic analysis of speech is for all practical purposes based on digital signal processing (59,60). The analog signal is converted to digital form for storage in a computer, often a microcomputer. Software programs enable the user to perform a variety of analyses and displays. Some of the functions are summarized below.

Waveform display and editing: A selected portion of the speech waveform can be displayed on a video monitor. Typically, cursors are then positioned to make further selections for editing, analysis, or playback. Figure 11c gives an example of a wave-

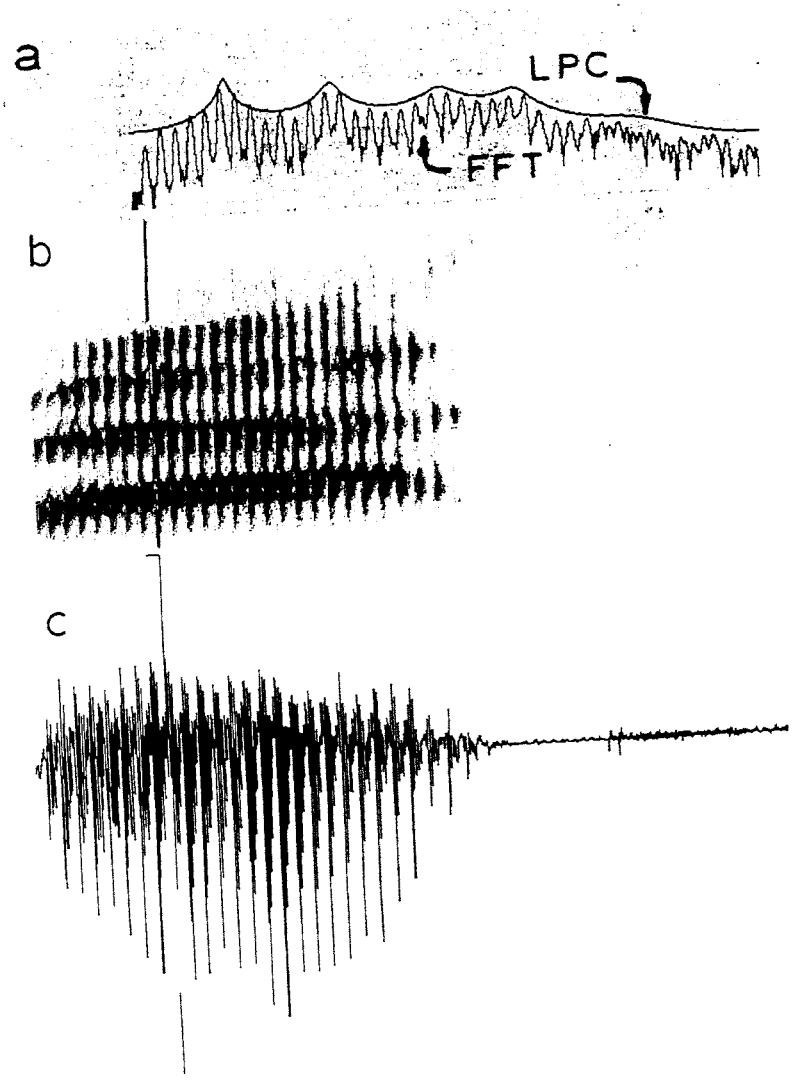

FIG. 11. Acoustic analysis of speech showing (a) spectral analyses, (b) spectrogram and (c) waveform. The waveform in (c) shows a cursor positioned to select a point for spectral analysis; (b) the spectrogram corresponds to the waveform in (a) and also shows the selected cursor position; and the FFT (fast Fourier transform) and LPC (linear predictive coding) spectra in (a) were determined for the cursor positions shown.

form display with a cursor positioned to mark a point in a vowel. Once the segment is marked, it can be played for listening or analyzed further as discussed below. In addition, manipulation and measurement of the waveform greatly aids the investigation of the temporal structure of speech, which is a rich source of information for segmental and prosodic aspects of speech (61–63).

Spectrogram: Although the analog spectrograph has all but disappeared from most laboratories, the spectrogram as an analysis format survives, now generated from digital files and displayed on a video monitor or printed with a hard copy device. Many speech analysis systems permit the waveform and spectrogram to be simultaneously displayed. Figure 11b shows the spectrogram that corresponds to the waveform in Fig. 11c. Another tool that may become increasingly used in speech analysis is the Wigner-Ville distribution, which, like the spectrogram, is a time-frequency representation. This distribution gives the advantage of high resolution at the expense of interference (cross) terms in the time-frequency spectrum. Time-frequency smoothing can be used to suppress the cross-terms, and, if the fundamental period is taken as the time resolution of the analysis, then the time-smoothed Wigner-Ville momentum spectrum is a "pitch spectrum" (64).

Spectral analysis: Almost all of the commercial speech analysis systems for microcomputers provide for fast Fourier transform (FFT) and linear predictive coding (LPC) spectra. Figure 11a shows an overlaid display in which important differences between the two analyses are readily seen. The segment selected for analysis is the vowel waveform shown in Fig. 11c. The FFT spectrum makes apparent the harmonics of the voicing source. The relative amplitudes of the harmonics as seen in the FFT spectrum reflect the combined source spectrum, the transfer function of the vocal tract, and the radiation characteristic. If the purpose of analysis is to determine formant frequency locations, then the user must infer the formant structure from the harmonic spectrum. This task is often uncertain, especially if the actual center frequency of the formant is not coincident with a harmonic. The uncertainty is proportional to the speaker's fundamental frequency; i.e., the error in formant frequency estimation is greater for voices with a high fundamental frequency. In contrast, the LPC analysis shows a spectral envelope that is ideally related only to the effects of formant shaping. The identification of formants is therefore easier, as the peaks in the spectrum presumably reflect formant structure. In practice, the LPC analysis does not always reveal formants as clearly as in the illustration. Particularly when formants are close together, as in the case of F1 and F2 for vowels [u] and [a], the analysis can fail to resolve the proximal formants. Another analysis now beginning to appear in microcomputer-based speech analysis systems is formant tracking, typically based on a LPC spectral analysis. This analysis provides an automatic tracking of formant frequencies. A particular advantage of this analysis is that the formant frequencies and bandwidths are available in a data file so that they can be analyzed statistically or used for purposes such as LPC resynthesis (described in a later section). Both FFT and LPC analyses are now fairly standard features of microcomputer-based systems and they are highly useful. However, the future may hold the development and refinement of still other approaches to spectral analysis. Hermansky (65) described a linear perceptual coding (PLP) that incorporates psychophysical properties of the human ear. An advantage to PLP analysis is that it may improve the correlation between acoustic analysis and auditory-perceptual judgments of the speech signal.

Vocal fundamental frequency extraction and related parameters of vocal function: Most speech analysis systems allow at least one technique for fo determination. The systems vary greatly in the algorithms used, speed of analysis, and vulnerability to various kinds of error (66,67). Many analysis systems also provide additional measures of vocal function. For example, CSpeech (68) calculates jitter, shimmer, and S/N ratio. The system developed by Nikolov et al. (69) extracts data for seven time-domain parameters and three frequency-domain parameters. In contrast, others have tried to develop single-value indices of voice or speech function. Harmegnies (70,71) used statistical indices of the degree of (dis)similarity between spectra. Frokjaer-Jensen and Prytz (72) also worked toward a single-value registration of spectral information. Dejonckere, Wieneke, and de Krom (73) used cepstral analyses. The cepstrum is a spectrum of a spectrum, i.e., the inverse Fourier transform of the power spectrum. Dejonckere et al. reported that the relative amplitude of the dominant rharmonic (highest peak in the cepstrum) correlated highly with voice quality judgments and seemed to be an effective objective synthesis of features of voice quality.

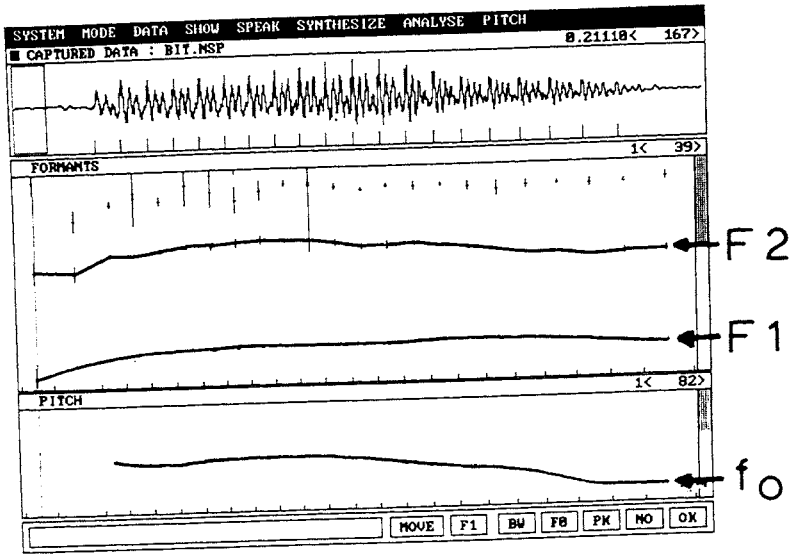

FIG. 12. Display of parameters derived from LPC analysis of the word *bit*. Note in particular the first two formants F1 and F2 and the intonation contour labeled fo.

In addition to the foregoing analyses, some systems offer additional capabilities that have great potential in understanding the acoustic aspects of voice and speech. One of these is LPC resynthesis, in which the coefficients extracted by LPC analysis can be selectively modified. Resynthesis is then performed on the modified coefficient matrix. An example is shown in Figs. 12 and 13 which illustrate modification of both formant pattern and fo contour. The original utterance *bit* [bIt] produced with a falling fo contour was resynthesized to form a new utterance *bat* [b ae t] having a rising fo contour. This resynthesis was performed on an IBM PC/AT microcomputer and an analysis system from Kay Elemetrics.

Resynthesis is a powerful tool that can be used to examine the effects of selected changes in acoustic features of a signal. It will enable the investigator or clinician to determine the effect of a selected acoustic change on the intelligibility or quality of a speaker's utterance. To date, resynthesis has been used largely with the speech of the deaf to determine which features are related to improvements in intelligibility. Maassen and Povel (74) concluded from their work on LPC resynthesis that improved intelligibility for the 10 subjects they examined would depend more on articulatory changes than on alterations of temporal structure or intonation.

More detailed discussions of speech analysis systems for microcomputers, including prices and ven-

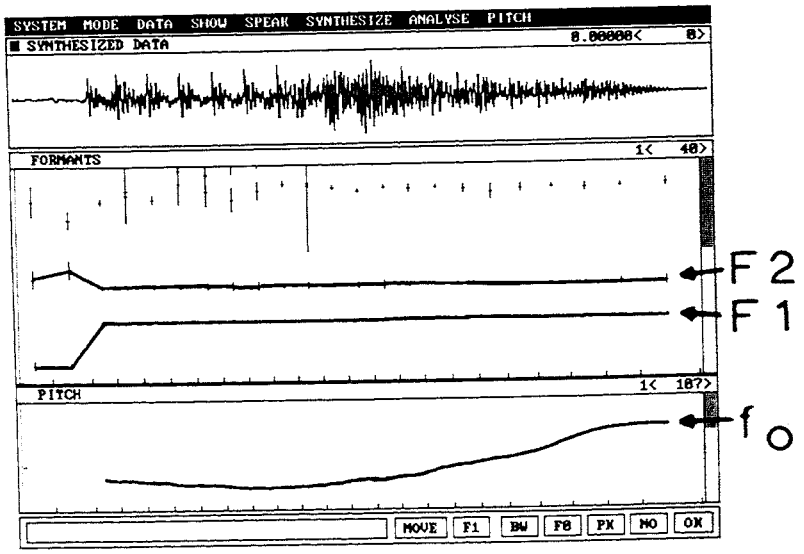

FIG. 13. Display of parameters for the word *bat*, resynthesized from the parameters shown in Fig. 12 for the word *bit*. The resynthesized sound has a different formant pattern (note the different F1 and F2 frequencies) and a different intonation (note the altered fo contour).

dor addresses, are available in Read, Buder and Kent (67,75).

ACOUSTIC MEASURES OF SPEECH AND VOICE

The acoustic signal of speech is rich in potential information, and a large number of measures have been proposed for its analysis (76–79). Sundberg (5) is a good source of acoustic information on singing. Some acoustic measures of speech were mentioned in the foregoing section. It is not possible to accomplish a comprehensive review of these measures in this paper. However, a relatively small set of measures have a high frequency of usage. These measures are summarized in Table 2 with respect to analysis techniques. The table entries indicate the suitability of a particular technique (e.g., waveform display, wide-band spectrogram, LPC spectrum) for a given measurement. When a question mark appears, it means that some relevant information conceivably could be produced with the technique but it is not conventional or efficient in this application. Other entries in the table are explained in the caption.

Tables 3 and 4 summarize major acoustic properties of vowels and consonants. Table 3 gives rules of thumb for the acoustic correlates of vowel production. Table 4 summarizes the acoustic correlates of major consonant classes.

The acoustic signal of speech contains much information on voice production and the resonance properties of the vocal tract. As indicated earlier, several analysis methods can be used to extract this information. No single method satisfies every purpose, but one can easily be overwhelmed by attempting to use several alternative analyses and an associated large number of acoustic measures. Moreover, choice of analysis rests on conceptual issues. Klingholz and Martin (80) described two concepts of how the voice functions as a signal generator. Each concept is associated with a class of analysis methods. One concept is deterministic, for example, attempts to identify acoustic-physiological relationships as they are expressed in the magnitudes and time records of acoustic variables. The other concept assumes that the speech wave is suitably represented as an ergodic random process, in which case the usual analysis effort is directed to description of the long-term distribution of an acoustic variable. The two approaches can be illustrated with the measure of jitter, or cycle-to-cycle variation in the fundamental period of the glottal waveform. A deterministic approach might seek to explain variations in jitter with respect to some set of physiological variables. Orlikoff (81) considered how jitter could be related to cardiovascular and neuromuscular variables. The former was regarded as a nonrandom influence and the latter as a random influence. Alternatively, if one assumed that jitter can be described solely as a random process, then a natural approach is to examine the long-term distribution of jitter for a voice sample.

The deterministic vs. random distinction is not dichotomous. Not only could there be a combination of deterministic and random influences, as suggested by Orlikoff (81), but there is an intriguing third possibility of explanation—chaos. Chaotic systems may seem upon superficial examination to be random processes, but if the initial conditions are carefully specified, the systems "settle" into predictable stable states. Chaos is being applied to many signals and natural phenomena that have long been regarded as random processes. For example, some recent analyses of the electroencephalogram seek to understand these electrical activity patterns of the brain in terms of "strange attractors" (82). The behavior of the vocal folds also may exhibit some patterns that can be accounted for by chaos. For example, different modes of vibration might be associated with different stable states. This is far too complex an idea to be summarized in any satisfactory way in this paper, but the point of this discussion is that scientists in many disciplines are turning to chaos to explain phenomena that were not satisfactorily explained by traditional notions of determinism and randomness. One observer has noted that, when the scientific accomplishments of the twentieth century are finally evaluated, just three will stand out: relativity, quantum mechanics and chaos.

PROSPECTS FOR AUTOMATED, MULTIDIMENSIONAL ANALYSIS

Finally, attention is given to some examples of acoustic analysis. The combination of acoustic theory, digital processing methods, and acoustic-phonetic measures, is very nearly at the stage of generating automatic (or nearly so) quantitative evaluations for utterances of variable length, using a personal computer and modest investment in associated hardware and software. This capability

TABLE 2. *Suitability of various analysis techniques for selected acoustic measures*

Measurement	Technique: Waveform	Technique: Envelope
Voice onset time	Y	Y
Segment duration	Y	Y
Formant frequency	N	N
Formant amplitude	N	N
Formant bandwidth	N	N
Mean fundamental frequency	Y	N
Fundamental frequency contour	Y	N
Consonant noise spectrum (burst or frication)	N	N
Harmonic spectrum	?	N
Voicing energy	Y	N
Noise in voiced signal	Y	N
Amplitude rise time	Y	Y
Jitter	Y (calc)	N
Shimmer	Y (calc)	N
Signal/noise ratio	Y (calc)	N

Measurement	Technique: Spectrogram: Wide-band	Technique: Spectrogram: Narrow-band
Voice onset time	Y	Y (poor)
Segment duration	Y	Y (poor)
Formant frequency	Y	Y
Formant amplitude	Y (poor)	Y (poor)
Formant bandwidth	Y (poor)	Y (poor)
Mean fundamental frequency	Y (poor)	Y
Fundamental frequency contour	Y (poor)	Y
Consonant noise spectrum (burst or frication)	Y	Y
Harmonic spectrum	NA (usually)	Y
Voicing energy	Y	Y
Noise components in voiced signal	Y	Y
Amplitude rise time	Y (poor)	Y (poor)
Jitter	Y (poor)	Y (poor)
Shimmer	Y (poor)	?
Signal/noise ratio	N	N

Measurement	Technique: FFT spectrum	Technique: LPC spectrum
Voice onset time	N	N
Segment duration	N	N
Formant frequency	Y (HP)	Y
Formant amplitude	Y (HP)	Y
Formant bandwidth	Y (HP)	Y
Mean fundamental frequency	N	N
Fundamental frequency contour	N	N
Consonant noise spectrum (burst or frication)	Y	Y
Harmonic spectrum	Y	N**
Voicing energy	Y	Y**
Noise in voiced signal	Y	Y
Amplitude rise time	N	N
Jitter	N	N
Shimmer	N	N
Signal/noise ratio	Y	Y

TABLE 2—(*Continued*)

Measurement	Technique: Cepstrum	Technique: Waterfall
Voice onset time	N	Y
Segment duration	N	Y
Formant frequency	N	Y
Formant amplitude	N	Y
Formant bandwidth	Y	Y (FFT)
Mean fundamental frequency	Y	Y (FFT)
Fundamental frequency contour	N	Y
Consonant noise spectrum (burst or frication)	N	Y (FFT)
Harmonic spectrum	Y	Y (FFT)
Voicing energy	Y	Y (FFT)
Noice in voiced signal	N	Y
Amplitude rise time	?	N
Jitter	?	N
Shimmer	?	?
Signal/noise ratio		

** Theoretically, an LPC analysis with a large number of coefficients can yield an harmonic spectrum like that produced by a FFT. Practically speaking, few commercially available systems allow such a large number of coefficients.

Y, yes, suitable; N, no, not suitable; ?, questionable; Y (calc), yes, with appropriate calculations; Y (poor), yes, but not a method of choice; NA, not applicable; Y (HP), yes, using harmonic pattern.

marks a profound advance over what was available only a decade or so ago.

An example of this capability is given in Figs. 14 and 15 for a short sentence "The potato stew is in the pot" spoken by a woman with dysarthria (a neurological speech impairment) resulting from amyotrophic lateral sclerosis, a degenerative neurological disease that typically results in severe dysarthria during its progression. The result shown in Fig. 14 is from a point relatively early in the disease when the woman had nearly normal speech. This illustration is a multi-parameter display in which quantitative analyses are automatically performed by a modified version of CSpeech (68). The results are plotted against time. The panels show, in descending order: (a) formant tracks determined by LPC; (b–e) coefficients derived from the fourth, third, second, and first moments of the spectral distribution; (f) rms amplitude envelope; and (g) fo as determined by a pitch determination algorithm.

Fig. 14 reveals that the woman is quite capable of making acoustic contrasts that are needed for intelligible speech. For example, the formant trajectories are substantial, showing that she was able to make significant adjustments in her vocal tract configuration. The coefficients from the spectral moments describe the overall shape of the spectrum as follows: M1, spectral mean or center of gravity;

TABLE 3. *Differences in selected acoustic measures between low vs. high vowels, front vs. back vowels, rounded vs. unrounded vowels, and nasal vs. nonnasal vowels*

Measure	Low-high difference		
Mean fo	Low vowel	<	High vowel
Intensity	Low vowel	>	High vowel
Duration	Low vowel	>	High vowel
F1 frequency	Low vowel	>	High vowel
Jitter	Low vowel	>	High vowel
	Front-back difference		
F2-F1 differene	Back vowel	<	Front vowel
	Tense-lax difference		
Duration	Tense vowel	>	Lax vowel
	Rounded-unrounded difference		
F1 + F2 + F3	Rounded vowel	<	Unrounded vowel
	Nasal-nonnasal difference		
Formant bandwidth	Nasal vowel	>	Nonnasal vowel
Intensity	Nasal vowel	<	Nonnasal vowel
F1 frequency	Nasal vowel	>	Nonnasal vowel
F2 + F3 frequency	Nasal vowel	<	Nonnasal vowel

M2, variance of energy about the mean; M3, skewness or tilt of the distribution; and M4, kurtosis (83). The changes in the first moment, M1, reflect significant changes in the spectrum, basically indicating that she produced fricatives (high M1 values) as well as vowels and sonorants (low M1 values).

Figure 15 shows the results of this multiparameter analysis for the same sentence produced by this woman several months later when she had a severe dysarthria. Note that the time scale of the figure has been adjusted because her speaking rate had slowed considerably. Her ability to make acoustic contrasts was markedly reduced at this point. Note in particular the flattened formant trajectories and the relatively unchanging first moment. These features can be related to the underlying physiology, especially atrophy of the tongue and a reduction of lingual motoneurons. The tongue eventually fails to move sufficiently to bring about large changes in formant pattern or to accomplish the constriction needed for turbulence noise.

The interpretations given so far are examples of a deterministic approach to analysis. Using the same data, one can take another perspective based on the long-term average spectrum. Fig. 16 shows the long-term power spectrum for the sentence seen in the multiparameter time history of Fig. 15. Distinct peaks in the spectrum correspond to the nearly unchanging formants apparent in Fig. 15.

The analyses presented in Figs. 14 and 15 go beyond the spectrogram in that they display quantitative parameters of speech and vocal function. The measures can be viewed in either a deterministic framework or a random ergodic framework. The spectrogram is a highly useful form of analysis but is not immediately quantitative in the sense of providing measures on speech and vocal function. Such measures are often tediously obtained by the user, working either from hard copy spectrograms or the monitor-displayed spectrogram. This is not to say that the spectrogram has lost its usefulness. CSpeech generates a spectrogram which is very important for evaluating the success of the LPC formant tracker and well as in aiding interpretation of features such as those that appear in Figs. 15 and 16.

The same analysis system, CSpeech, can be used to perform perturbation analyses, such as calculations of jitter, shimmer, and signal-to-noise ratio (Pinto and Titze (84) summarize these perturbation measures). Figure 17 shows a voice analysis for a sustained vowel produced by a man who has dysarthria associated with amyotrophic lateral sclerosis. The figure shows, from top to bottom, the waveform envelope, the fundamental frequency contour, the jitter waveform and the shimmer wave-

TABLE 4. *Acoustic properties of major classes of consonants*

	Consonant Class					
Property	F	S	A	N	L	G
Prolonged noise	Yes	No	Yes	No	No	No
Noise burst (brief noise)	No	Yes	No	No	No	No
Rapid rise of noise energy	No	Yes	Yes	No	No	No
Sonorant formant pattern	No	No	No	Yes	Yes	Yes
Nasal murmur	No	No	No	Yes	No	No
Stop gap (silent interval)	No	Yes	Yes	No	No	No
Voiced/voiceless cognates	Yes	Yes	Yes	No	No	No

F, fricative; S, stop; A, affricate; N, nasal; L, liquid; G, glide.

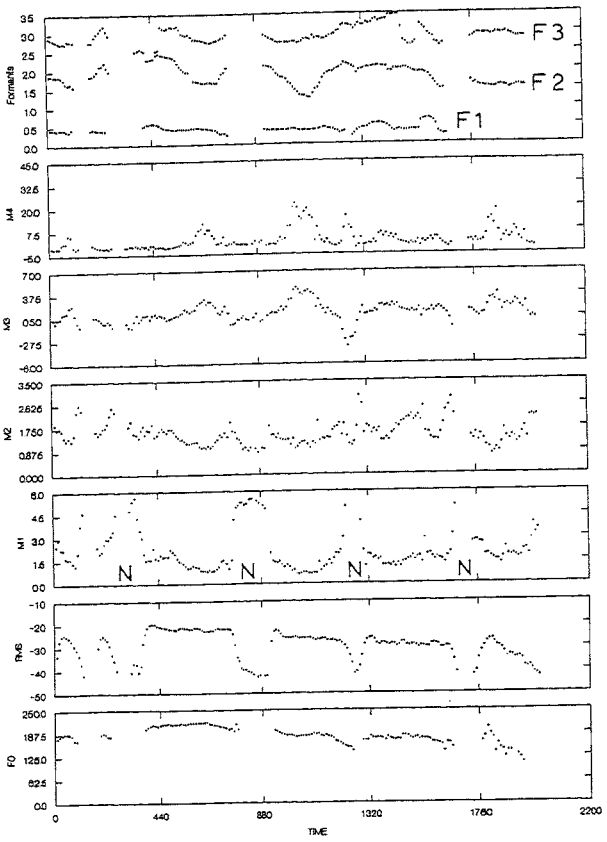

FIG. 14. Multiparameter acoustic analysis of the sentence, "The potato stew is in the pot," spoken by a woman who has an early stage of amyotrophic lateral sclerosis. See text for description. Note especially the changing formants (F1, F2, F3) and the increase in the M1 coefficient for noise (N) segments. (The formant tracking algorithm did not follow F1 to the end of the utterance.)

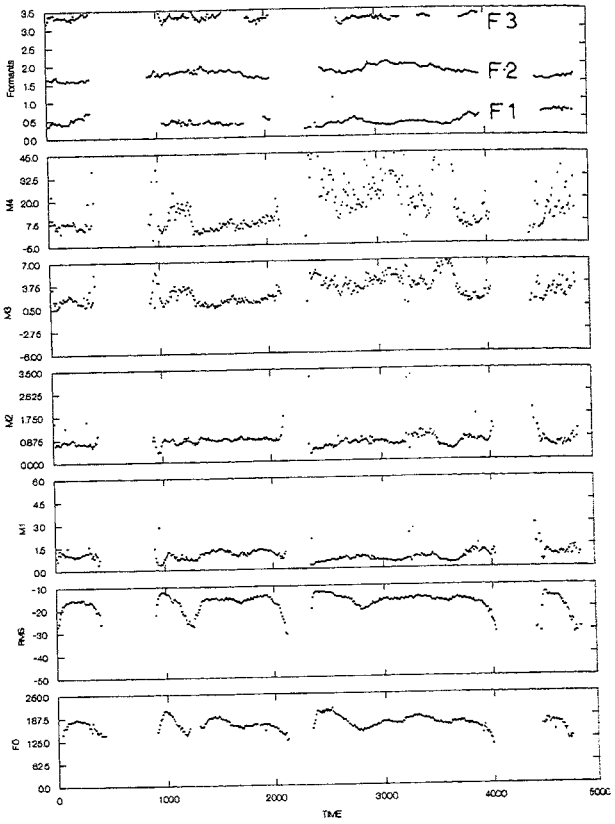

FIG. 15. Multiparameter acoustic analysis of the sentence, "The potato stew is in the pot," spoken by a woman with advanced amyotrophic lateral sclerosis. She is severely dysarthric in her production. Compare with Fig. 14 and see text for description. Note especially the relatively flat formant trajectories and the nearly constant M1 coefficient.

form. Usually, jitter and shimmer values are reported numerically, but we find that plotting the perturbation values helps to identify temporal variations in vocal function, e.g. episodes of marked instability.

SUMMARY

The rapidly evolving field of speech acoustics is making available to the scientist, clinician, teacher, and hobbyist a powerful but reasonably economical capability for acoustic analyses that, as recently as a few years ago, could be accomplished only with a mainframe computer and expensive hardware-software systems. The prospective user of a contemporary system for acoustic analysis of speech and voice can develop a multifunction laboratory around a microcomputer. The study and application of vocal tract acoustics has entered a new era of data collection and analysis. This era is based on a foundation of linear source-filter theory, quantitative analyses, and efficient methods of digital signal processing that can be implemented on microcomputers. A welcome feature of modern analysis sys-

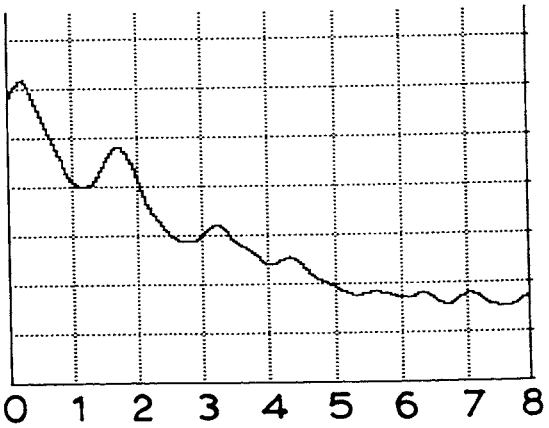

FIG. 16. Long-term power spectrum for the sentence displayed in Fig. 15. Note prominence of spectral peaks corresponding to the relatively constant formants. Frequency in kHz is scaled on the abscissa. Ordinate is relative amplitude.

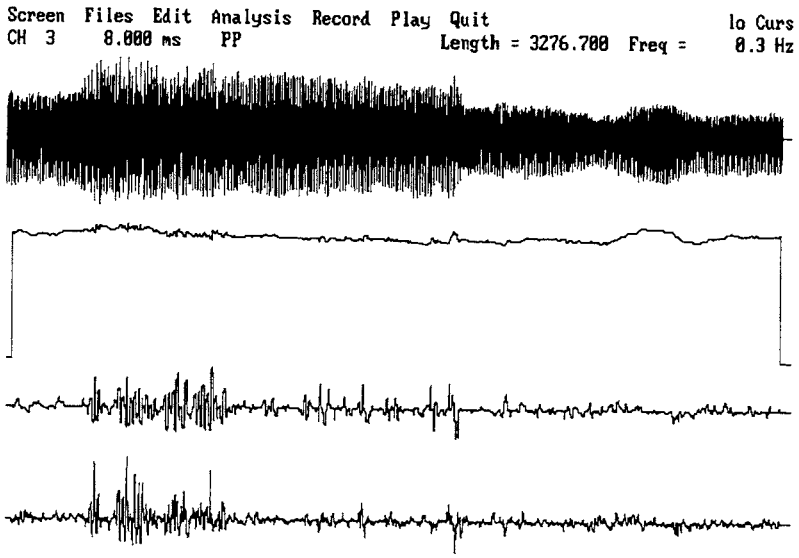

FIG. 17. Acoustic analysis of vowel phonation from a man with dysarthria associated with amyotrophic lateral sclerosis. The analyses displayed are the amplitude envelope, fundamental frequency, jitter, and shimmer. The latter are measures of perturbations in the laryngeal waveform.

tems is that they can provide measures of both voice and vocal tract function. The correct interpretation of this information requires a knowledge of acoustic theory and analysis algorithms. This paper has summarized basic speech acoustics necessary for such interpretations.

Acknowledgment: This work was supported in part by NIH research grants DC00319 and DC00490 from the National Institute on Deafness and Other Communication Disorders. Professor Paul Milenkovic and Dr. Eugene Buder contributed to the analyses described in this article. Dr. Ronald Scherer offered several helpful comments on an earlier version of the manuscript; his suggestions are gratefully acknowledged.

REFERENCES

1. Fant G. *Acoustic theory of speech production*, 2nd Ed. The Hague: Mouton, 1960.
2. Teager HM, Teager SM. Evidence for nonlinear sound production mechanisms in the vocal tract. In Hardcastle WJ, Marchal A, (eds). *Speech production and speech modelling*. Dordrecht, Netherlands: Kluwer, 1990:241–61.
3. Fant G. Glottal flow: models and interaction. *J Phonetics* 1986;14:393–9.
4. Sundberg J. *The science of the singing voice*. DeKalb, IL: Northern Illinois University Press, 1987.
5. Sondhi MM. Resonances of a bent vocal tract. *J Acoust Soc Am* 1986;79:1113–6.
6. Fujimura O. Methods and goals of speech production research. *Lang Speech* 1990;33:195–258.
7. Fujimura O, Lindqvist J. Sweep-tone measurements of vocal-tract characteristics. *J Acoust Soc Am* 1971;49:541–58.
8. Klatt DH. Software for a cascade/parallel formant synthesizer. *J Acoust Soc Am* 1980;67:979–95.
9. Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am* 1990;87:820–57.
10. Markel J, Gray A. *Linear prediction of speech*. Berlin: Springer-Verlag, 1976.
11. Noll AM. Short-time spectrum and "cepstrum" techniques for vocal pitch detection. *J Acoust Soc Am* 1964;36:296–302.
12. Strange W. Evolving theories of vowel perception. *J Acoust Soc Am* 1987;85:2081–7.
13. Peterson GE, Barney HE. Control methods used in a study of vowels. *J Acoust Soc Am* 1952;24:175–84.
14. Bergem DR, van Pols LCW, Koopmans-van Beinum FJ. Perceptual normalization of the vowels of a man and a child in various contexts. *Speech Commun* 1988;7:1–20.
15. Fant G. Non-uniform vowel normalization. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm. *Quart Prog Status Rep* 1975;2–3:1–9.
16. Verbrugge RR, Strange W, Shankweiler DP, Edman TR. What information enables a listener to map a talker's vowel space? *J Acoust Soc Am* 1976;60:198–212.
17. Disner SF. Evaluation of vowel normalization procedures. *J Acoust Soc Am* 1980;67:253–61.
18. Syrdal AK, Gopal HS. A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J Acoust Soc Am* 1986;79:1086–1100.
19. Stevens KN, House AS. Development of a quantitative description of vowel articulation. *J Acoust Soc Am* 1955;27:484–93.
20. Badin P, Perrier P, Boe L-J, Abry C. Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *J Acoust Soc Am* 1990;87:1290–1300.
21. Liljencrants J. Fourier series description of the tongue profile. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm. *Quart Prog Status Rep* 1971;4:9–18.
22. Harshman R, Ladefoged P, Goldstein L. Factor analysis of tongue shapes. *J Acoust Soc Am* 1977;62:693–707.
23. Kiritani S. Articulatory studies by the X-ray microbeam system. In: Sawashima M, Cooper FS (eds). *Dynamic aspects of speech production*. Tokyo: University of Tokyo, 1977:1711–90.
24. Maeda S. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle WJ, Marchal A, (eds). *Speech production and speech modelling*. Dordrecht, Netherlands: Kluwer, 1990:131–49.
25. Lindblom BEF, Sundberg JEF. Acoustical consequences of

lip, tongue, jaw and larynx movement. *J Acoust Soc Am* 1971;50:1166–79.
26. Mermelstein P. Articulatory model for the study of speech production. *J Acoust Soc Am* 1973;53:1070–82.
27. Rubin PE, Baer T, Mermelstein P. An articulatory synthesizer for perceptual research. *J Acoust Soc Am* 1981;70: 321–8.
28. Stevens KN. On the quantal nature of speech. *J Phonetics* 1989;17:3–45.
29. Wood S. A radiographic analysis of constriction locations for vowels. *J Phonetics* 1979;7:25–43.
30. Perkell JS, Nelson WL. Variability in production of the vowels /i/ and /a/. *J Acoust Soc Am* 1985;77:1889–95.
31. Carre R, Mrayati M. Articulatory-acoustic-phonetic relations and modelling, regions and modes. In: Hardcastle WJ, Marchal A (eds). *Speech production and speech modelling*. Dordrecht, Netherlands: Kluwer, 1990:211–40.
32. Fujimura O. Analysis of nasal consonants. *J Acoust Soc Am* 1962;34:1865–75.
33. Kurowski K, Blumstein SE. Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *J Acoust Soc Am* 1984;76:383–90.
34. Repp BH, Svastikula K. Perception of the [m]-[n] distinction in VC syllables. *J Acoust Soc Am* 1988;83:237–47.
35. Kent RD, Read CW. *The Acoustic Analysis of Speech*. San Diego, CA: Singular Publishing Group, in press.
36. Shadle CH. Articulatory-acoustic relationships in fricative consonants. In: Hardcastle WJ, Marchal A (eds). *Speech production and speech modelling*. Dordrecht, Netherlands: Kluwer, 1990:187–209.
37. Hughes GW, Halle M. Spectral properties of fricative consonants. *J Acoust Soc Am* 1956;28:303–10.
38. Strevens P. Spectra of fricative noise in human speech. *Lang Speech* 1960;3:32–49.
39. Stevens KN. Airflow and turbulence noise for fricative and stop consonants: static considerations. *J Acoust Soc Am* 1971;50:1180–92.
40. Pentz A, Gilbert HR, Zawadski P. Spectral properties of fricative consonants in children. *J Acoust Soc Am* 1979;66: 1891–92.
41. Bauer HR, Kent RD. Acoustic analysis of infant fricative and trill vocalizations. *J Acoust Soc Am* 1986;81:501–11.
42. Delattre P, Liberman AM, Cooper FS. Acoustic loci and transitional cues for consonants. *J Acoust Soc Am* 1955;27: 769–74.
43. Halle M, Hughes GW, Radley JP. Acoustic properties of stop consonants. *J Acoust Soc Am* 1957;29:107–16.
44. Heinz JM and Stevens KN. On the properties of voiceless stop consonants. *J Acoust Soc Am* 1961;33:589–96.
45. Liberman AM, Cooper FS, Shankweiler DS, Studdert-Kennedy M. Perception of the speech code. *Psychol Rev* 1967;74:431–61.
46. Repp BH, Liberman AM, Eccardt T, Pesetsky D. Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *J Exper Psychol Hum Percept Perform* 1978;4: 621–37.
47. Kewley-Port D. Time-varying features as correlates of place of articulation in stop consonants. *J Acoust Soc Am* 1983; 73:322–35.
48. Kewley-Port D. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *J Acoust Soc Am* 1983;72:379–89.
49. Kewley-Port D, Pisoni DB, Studdert-Kennedy M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *J Acoust Soc Am* 1983;73: 1779–93.
50. Klatt DH. Voice onset time, frication and aspiration in word-initial consonant clusters. *J Speech Hear Res* 1975;18: 686–706.
51. Howell P, Rosen S. Production and perception of rise time in the voiceless affricate/fricative distinction. *J Acoust Soc Am* 1983;73:976–84.
52. Nolan F. *The phonetic bases of speaker recognition*. Cambridge UK: Cambridge University Press, 1983.
53. Liberman AM, Delattre PD, Cooper FS, Gerstman LJ. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J Exper Psychol* 1954;52:127–37.
54. O'Connor JD, Gerstman LJ, Liberman AM, Delattre PC, Cooper FS. Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word* 1957;13:24–43.
55. Lehiste I, Peterson GE. Transitions, glides, and diphthongs. *J Acoust Soc Am* 1961;33:268–77.
56. Dalston R. Acoustic characteristics of English /w,r,l/ spoken correctly by young children and adults. *J Acoust Soc Am* 1975;57:462–69.
57. Fant G. The relations between area functions and the acoustic signal. *Phonetica* 1980;37:55–86.
58. Baken R, Daniloff R (eds). *Readings in clinical spectrography of speech*. San Diego: Singular Publishing Group, 1990.
59. Fallside F, Woods WA. *Computer speech processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.
60. Pohlman KC. *Principles of digital audio*. Indianapolis, IN: Howard W. Sams, 1985.
61. Umeda N. Vowel duration in English. *J Acoust Soc Am* 1975;58:434–45.
62. Klatt DH. Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *J Acoust Soc Am* 1976; 59:1208–21.
63. Crystal TH, House AS. Segmental durations in connected speech signals: preliminary results. *J Acoust Soc Am* 1982; 72:705–16.
64. Deliyski D, Zielinski T. Objective diagnosis of laryngeal pathology using the Wigner-Ville distribution. In: Proceedings of the 12th Signal Processing Symposium GRETSI'89, Juan-les-Pins, France: August 16, 1989.
65. Hermansky H. Perceptual linear predictive (PLP) analysis of speech. *J Acoust Soc Am* 1990;87:1738–52.
66. Hess W. *Pitch determination of speech signals*. New York: Springer-Verlag, 1983.
67. Read C, Buder EH, Kent RD. Speech analysis systems: an evaluation. *J Speech Hear Res* 1992;35:314–32.
68. CSpeech, a software program for speech analysis. Available from Paul Milenkovic, Electrical & Computer Engineering, University of Wisconsin-Madison.
69. Nikolov Z, Diliyski D, Drumeva L, Boyanov B. Computer system for analysis of pathological voices. Paper given at XXI Intern Assoc Logopedics Phoniatrics, Prague, August 1989.
70. Harmegnies B. Contribution a la caracterisation acoustique de la qualite vocale. Analyses plurielles de spectres moyens a long terme de parole. Ph.D. Dissertation, Universite de Mons, France, 1988.
71. Harmegnies B. SDDD, a new dissimilarity index for the comparison of speech spectra (Letter). *Pattern Recognit* 1988;8:153–8.
72. Frokjaer-Jensen B, Prytz S. Registration of voice quality. *Bruel Kjaer Technol Rev* 1976;3:3–17.
73. Dejonckere PH, Wieneke GH, de Krom G. Cepstra of normal and pathological voices in relation to acoustical and perceptual data. Paper presented at the 1st Intern. Clin. Phonetics & Linguistics Assoc Symposium (Cardiff, Wales UK), March 25–26, 1991.
74. Maassen B, Povel DJ. The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. *J Acoust Soc Am* 1985;78:877–86.
75. Read C, Buder EH, Kent RD. Speech analysis systems: a survey. *J Speech Hear Res* 1990;33:363–74.
76. Shoup JE, Pfeifer LL. Acoustic characteristics of speech

sounds. In: Lass NJ (ed). *Contemporary/Issues in experimental phonetics* New York: Academic Press, 1976;171–224.
77. Fry DB. *The physics of speech.* Cambridge, U.K.: Cambridge University Press, 1979.
78. Picket JM. *The Sounds of Speech Communication.* Baltimore: University Park Press, 1980.
79. Kent RD, Read WC. *Acoustic analysis of speech.* San Diego: Singular Publishing Group, 1992:238 pp.
80. Klingholz F, Martin F. Distribution of the amplitude in the pathologic voice signal. *Folia Phoniatr* 1989;41:23–9.
81. Orlikoff RF. Vocal jitter at different fundamental frequencies: a cardiovascular-neuromuscular explanation. *J Voice* 1989;2:104–12.
82. Basar E. *Chaos in Brain Function.* Berlin: Springer-Verlag, 1990.
83. Forrest K, Weismer G, Milenkovic P, Dougall RN. Statistical analysis of word-initial voiceless obstruents: preliminary data. *J Acoust Soc Am* 1988;84:115–23.
84. Pinto NB, Titze IR. Unification of perturbation measures in speech signals. *J Acoust Soc Am* 1990;87:1278–89.