Acoustic Characteristics of Normal and Pathological Voices

Steven B. Davis

ABSTRACT

Recent studies have suggested that acoustic analysis of the pathological voice is a viable technique for early detection of laryngeal pathology or for clinical assessment of vocal improvement following voice therapy. Acoustic parameters may be extracted directly from speech or throat contact signals or indirectly from glottal or residue inverse filtered signals. These parameters characterize voice quality differences and may be used to discriminate between normal and pathological subjects. Such parameters may measure temporal variation, such as average pitch period or amplitude perturbation, or spectral variation, such as energy differences among frequency bands. This chapter discusses methods for the acoustic analysis of voices affected by laryngeal pathology, and procedures for determining acoustic parameters applicable to screening and clinical assessment. These methods use digital computer techniques for voice analysis to extract acoustic measures of vocal function from the speech signal. A brief review provides information on some auditory and visual methods for diagnosing laryngeal pathology. Then the vocal fold movement is related to the acoustic output on which the acoustic methods are based, and the theoretical bases and results of several methods are compared. Finally, acoustic parameters and representative waveforms based on inverse filtered speech are used in a Voice Profile to assess early cases of pathology and to monitor progress during voice therapy.

INTRODUCTION

In recent years, researchers in such fields as laryngology and speech science have become increasingly interested in the acoustic characteristics of normal and pathological voices (Murry, 1975; Davis, 1975, 1976b; Hiki, Imaizumi, Hirano, Matsushita and Kakita, 1976). One reason for this trend is


+Also Speech Communications Research Laboratory, Inc., Santa Barbara, California.

Acknowledgement: The preparation of this chapter was supported by National Institutes of Health Grants NS 13309-01A1 to Speech Communications Research Laboratory, Inc. and NS 13870 to Haskins Laboratories and the Voice Foundation, New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]
that acoustic methods have the potential to provide quantitative techniques for the clinical assessment of laryngeal function. The desire for an objective method for analysis of the pathological voice was expressed by the 1973 Conference on Early Detection of Laryngeal Pathology (Moore, 1973, p. 6):

"The otologist and audiologist can employ standardized audiometric tests to evaluate hearing, the cardiologist has access to electrocardiograms to evaluate heart function, but the laryngologist and speech pathologist have no comparable aids that can be used in the clinical setting."

There are several methods currently used in laryngeal research and diagnosis, for example, laryngoscopy, stroboscopy, thermography, electromyography, pneumotachography, glottography and acoustic analysis, but acoustic analysis appears to have an advantage for routine clinical evaluation of laryngeal function, for example, during a program of voice therapy, because of its nonintrusive nature and its potential for providing quantitative data with reasonable expenditures of analysis time.

In addition to its potential value in the clinical assessment of laryngeal function, a sensitive automatic acoustic technique could be used to screen individuals for early cases of laryngeal pathology. Moore (1973) has indicated that none of the other techniques are useful for screening, and are only applied if an individual specifically seeks aid. The development of portable instrumentation for acoustic analysis would lead to programs similar to audiometric testing in schools, industry, etc. Such instrumentation could have large benefits in terms of overall health maintenance.

This chapter discusses methods for the acoustic analysis of voices affected by laryngeal pathology and procedures for determining acoustic parameters applicable to screening and clinical assessment. These methods use digital computer techniques for voice analysis (developed originally to increase the efficiency of speech transmission systems) to extract acoustic measures of vocal function from the speech signal. A brief review provides information on some auditory and visual methods for diagnosing laryngeal pathology. Then the vocal fold movement is related to the acoustic output on which the acoustic methods are based, and the theoretical bases and results of several methods are compared, indicating the difficulties requiring future research.

**AUDITORY AND VISUAL METHODS**

Historically, physicians have relied on two basic techniques in the diagnosis of pathological conditions of the larynx: 1) listening to the voice, and 2) viewing the larynx with a mirror or laryngoscope. Since laryngeal diseases are often accompanied by voice quality changes, simple listening tests sometimes give useful information. The principal criticisms of listening tests are their subjectivity (that is, even experienced laryngologists may offer different diagnoses for the same patient), and the related problem of the lack of quantitative standards.

Visual observations allow a physician to substantiate auditory evaluations. In indirect laryngoscopy, the larynx is viewed via a mirror inserted
into the back of the mouth. The gross structure and movements of the vocal folds are observed; however, the amount of detail made available by indirect laryngoscopy is limited because the field of view is small and the distance from the larynx is relatively long. Pathologies beneath the vocal folds frequently can be overlooked because only the superior surfaces of the larynx may be visualized from above. Also, the rapid vibratory motions of the vocal folds cannot be observed with indirect laryngoscopy. However, high quality photographic records of the laryngeal image exposed by advanced laryngoscopes using fiberoptics (Sawashima and Hirose, 1968; Gould, 1973) have enhanced the clinical value of laryngoscopy.

In direct laryngoscopy a viewing tube is inserted directly into the larynx. Direct laryngoscopy is not used on a wide scale because it is a medical procedure requiring anaesthesia, it disturbs the positioning and function of the articulatory structures, and it is uncomfortable for the patient. Consequently, its application usually is confined to diagnosis and verification during the later stages of a laryngeal disease and to surgical situations (Koike, 1976).

Stroboscopic laryngoscopy combines a laryngeal mirror and a high flash-rate stroboscope to give either a stationary or slowly moving image of the vocal folds (Moore, 1938; Schönhärl, 1960; van den Berg, 1962). However, the image is a composite of many cycles of the vibration, and fine details of the individual periods are not observed.

The detailed movement of the vocal folds during individual periods can be seen, however, if a high-speed motion picture camera, a special light source and a laryngeal mirror are used (Farnsworth, 1940; Koike and Takahashi, 1971). Compared to stroboscopic laryngoscopy, this high-speed technique captures all vibratory behavior in full detail. Studies based on high-speed motion pictures have demonstrated various vibratory patterns in patients with laryngeal pathology (Timcke, von Leden and Moore, 1958, 1959; von Leden, Moore and Timcke, 1960). In one study (Moore, 1968), high-speed motion pictures of pathological vocal folds were synchronized with acoustic recordings of the voice. Moore demonstrated complex and irregular vibratory patterns resulting in complicated changes in glottal width. He also noted the independence of each fold as a vibrator and suggested that this independence should be considered as a cause of hoarseness. This independence was confirmed by Koike and Hirano (1973).

High-speed motion picture analysis is an important tool in basic voice and speech research, but there are several limitations for large-scale clinical applications. The most significant limitation is the time-consuming task of frame-by-frame data analysis. Even with the aid of digital computer programs designed to simplify the measurement process (Ramsey, 1964; Soron, 1967; Koike and Takahashi, 1971; Hayden and Koike, 1972; Tanabe, Kitajima, Gould and Lamblase, 1975; Hildebrand1), it is not a trivial task to record the

---

glottal widths or areas for a large number of frames. One preliminary study (Davis, 1976a) applied digital image processing to automatically extract the glottal area information from successive picture frames, but more research is needed before the method can become clinically useful. Nevertheless, the information obtained from high-speed motion pictures that are synchronized with the voice provides a basis for gaining a better understanding of the relationship between the acoustic signal and the physiological function of the larynx. Some of the parameters that can be measured from successive high-speed motion picture frames are the open quotient (OQ), which is the ratio of the open time of the glottis to the total time of one vibratory cycle, and the speed quotient (SQ), which is the ratio of the time of vocal fold abductory movement to the time of adductory movement. These parameters are important for assessing the vibratory behavior of the vocal folds (Timcke, von Leden and Moore, 1958, 1959; von Leden, Moore and Timcke, 1960; Hildebrand, see footnote 1).

**ACOUSTIC SYMPTOMS OF LARYNGEAL PATHOLOGY**

Acoustic analysis of the voice is more objective than auditory methods for screening or voice therapy assessment (Koike, 1976). The validity of the acoustic approach, however, rests on the complex relationship between the physiological source function and the concomitant speech signal. A laryngeal pathology, such as tumors or paralysis, generally produces asymmetrical changes in the mass, elasticity and tension of the vocal folds, leading to deviant vibration. Also, weakness or paralysis of respiratory muscles may cause insufficient subglottal pressure, thus changing the aerodynamic forces acting on the vocal folds and hence their vibratory pattern. The subglottal airstream is modulated by this unbalanced vocal fold movement. Irregular air pulses emerge from the larynx, propagate through the pharynx and oral and nasal cavities, and radiate from the mouth and nose. The resultant acoustic signal is thus affected by a physiological disturbance in the larynx, and the acoustic signal may be used to measure the disturbance.

The primary acoustic symptoms of laryngeal pathology are a change in fundamental frequency, voice intensity or voice quality. These symptoms are indicative of a multitude of organic diseases and functional disturbances (Zemlin, 1968; Moore, 1971), and the nature of these symptoms will vary for each patient and at each stage of pathological involvement.

**Fundamental Frequency**

The fundamental frequency of a voiced sound is a function of the mass, elasticity, compliance and length of the vocal folds. It also depends somewhat on the subglottal pressure and the configuration (acoustic load) of the vocal tract. An assessment of whether a patient has adequate frequency regulation usually involves a judgment by a trained listener as to whether the fundamental frequency is too high or too low when compared to voices of persons of similar age, sex and body size. If the fundamental frequency is judged to be too high, the voice may sound "shrill" or "screechy." In functional disorders involving high fundamental frequency, the vocal folds tend to become abused at the point of maximum displacement. Vocal abuse may produce laryngitis, lead to the development of nodules, or worsen an existing pathology. Organic causes for high fundamental frequency include a laryngeal

136
web, asymmetrical structures or failure of a male larynx to develop to a normal size. A fundamental frequency which is judged to be too low may sound "harsh," "hoarse," "husky" or "rough." Low fundamental frequency stemming from functional vocal abuse may lead to contact ulcers, although the precise etiology of contact ulcers is unclear. Organic causes may be virilization, tumors or other enlargements that increase the mass of the folds, or nerve paralysis that decreases the elasticity and compliance of one or both folds (Luchsinger and Arnold, 1965).

Vocal Intensity

Vocal intensity is a monotonically increasing function of the SQ and the air flow through the glottis, and it is a monotonically decreasing function of the OQ (Timcke, von Leden and Moore, 1958). Excessive vocal intensity usually is functional in origin. If it is also accompanied by high fundamental frequency, the voice may sound "shrill" or "screechy," and if it is also accompanied by low fundamental frequency, the voice may sound "hoarse." When excessive vocal intensity is coupled with excessively high or low fundamental frequency, the severity of a pathology may increase. Weak voices generally have organic causes, and the etiology may be attributed to insufficient subglottal pressure caused by paralysis of the respiratory muscles, or to poor vocal fold movement caused by muscle weakness or laryngeal paralysis (Luchsinger and Arnold, 1965).

Vocal Quality

A degradation in voice quality, generally categorized as "hoarseness," is often the first and sometimes the sole symptom of laryngeal disease. This change, because of its initial presence, drew the early attention of several researchers (Jackson and Jackson, 1937; Frank, 1940; Arnold, 1955; Palmer, 1959; Bowler, 1964). However, these studies were perceptual, and there was a multitude of concepts and terms. One review (Perkins, 1971) compared nine studies that assess quality defects and listed twenty-seven terms used to describe those defects. Only "hoarseness" and "nasality" appear in all studies, and only ten other terms appear in more than one study. Of these ten terms, six are common to four studies: "breathy," "harsh," "strident," "denasal," "husky," and "metallic." Other terms are "screechy," "guttural," "throaty," "strained," "shrill" and "intense." This review demonstrates that there is little agreement among researchers on describing voice quality, and there is a proliferation of descriptive terms.

This complex terminology led some researchers to attempt to identify the factors that are involved in listener judgments of pathological voice quality (Isshiki, 1966; Isshiki, Okamura, Tanabe and Morimoto, 1969; Takahashi and Kolke, 1975; Fritzell, Hammarberg and Wedin, 1977). Using a technique based on the semantic differential (Osgood, Suci and Tannenbaum, 1957), Isshiki (1966) suggested that the factors that operate in listener judgments are multidimensional, and he identified four major independent axes corresponding approximately to "roughness," "breathiness," "lack of power," and "normalcy." Using principal components analysis, Fritzell, Hammarberg and Wedin (1977) identified five factors as "stable-unstable," "breathy-overtight," "hypohyperkinetic," "light-course" and "head-chest register."
In summary, vocal quality is a difficult parameter to assess, and unlike fundamental frequency or vocal intensity, no reliable physiological descriptions or measurements have been established. In general, vocal quality is determined on the basis of vocal fold vibration and vocal tract resonance (Luchsinger and Arnold, 1965). Furthermore, asymmetrical vibrations are a typical indication of a vocal quality defect. Such variations affect the fundamental frequency, SQ and OQ, and may be attributed to changes in the mass, elasticity, compliance or length of one or both folds (von Leden, Moore and Timcke, 1960). Vocal phonatory defects range from "aphonia" and "breathiness" to "hoarseness" and "rough hoarseness," and the organic causes include paralysis, weak muscles, extraneous masses, excessive mucous and loss of tissue (Luchsinger and Arnold, 1965).

ACOUSTIC TECHNIQUES FOR VOICE ANALYSIS

Acoustic techniques for voice analysis are based on the source of the signal and the method of analysis. One source is direct signals, for example, radiated sound pressure and throat contact signals. The other source is indirect signals, for example, glottal or residue signals derived using inverse filtering techniques. For each of these sources, the signals may be analyzed in the time domain, for example, to determine mean fundamental frequency or mean perturbation, or in the frequency domain, for example, to determine long-term average spectral slope or energy distribution. The following sections will discuss the advantages and limitations of each signal source and analysis method.

Direct Signals

The radiated sound pressure waveform is the most readily available signal for acoustic analysis, but its usefulness for assessing laryngeal function is limited. The production of a steady vowel sound is controlled by the glottal source, which will be affected by laryngeal pathology, and the supraglottal structure, whose resonance characteristics will presumably be unaffected by laryngeal pathology. Therefore, acoustic measures of a laryngeal disorder from a sound pressure waveform are affected by a normal supraglottal structure. The effects of the supraglottal structure do not significantly hinder the detection or assessment of severe laryngeal disorders. However, at an early stage of pathology, or at a late stage of recovery, the supraglottal structure probably masks some of the important acoustic attributes of the pathology. A throat contact microphone is sometimes used to avoid supraglottal effects (Koike, 1969), but information from throat contact signals is limited by the low-pass filtering action of intervening tissues, and also because the throat signal is sensitive to microphone placement.

Indirect Signals

A voiced sound such as a sustained vowel may be simply modeled as the sound pressure waveform resulting from the excitation by a periodic source (corresponding to the glottis) of an acoustic tube (corresponding to the vocal tract and lips). The technique of inverse filtering applied to the sound pressure waveform can remove the effects of the acoustic tube, and the resulting signal approximates the periodic source. If the supraglottal structure is relatively unaffected by laryngeal pathology, and the source of a
voice change is the glottis, then this periodic source approximation contains sufficient information to analyze the acoustic effects of the pathology. Measures based on an inverse-filtered speech signal are not affected by the supraglottal structure, and are potentially more informative than measures based on an unfiltered speech signal.

There are two inverse filtering methods that are used to obtain acoustic information about the glottal source. The first method is glottal inverse filtering. In this procedure, the inverse of the lip radiation and vocal tract spectral contributions is used to estimate the glottal volume velocity waveform as a function of time (Miller, 1959; Holmes, 1962; Lindqvist, 1965; Takasugi and Suzuki, 1970; Rothenberg, 1973). The theoretical starting point for glottal inverse filtering is the linear voiced speech production model (Fant, 1960; Flanagan, 1972). Early glottal inverse filtering studies used analog techniques, and usually only eliminated the first and second formants. Later studies employed digital computer techniques to process the speech signal in an interactive manner. The formant frequencies and bandwidths were estimated and used to adjust a glottal inverse filter until the expected waveform (roughly triangular in shape followed by a zero portion) appeared (Hiki et al., 1976). However, the overall appearance of the resulting waveform generally did not satisfy intuitive concepts of vocal fold closure. The linear predictive technique of pitch-synchronous glottal inverse filtering produces acceptable waveforms without estimation of the formants, but the point of glottal closure for each pitch period must be located by visual inspection (Wong and Markel, 1976).

Another method of glottal inverse filtering eliminates the vocal tract resonance by having a subject speak into a fairly long reflectionless tube (approximately 2.5 cm x 100 cm) (Sondhi, 1975). The resulting waveform is closely correlated with the glottal volume velocity waveform. The method is subjective, however, since each speaker requires a different tube (matched to their vocal tract) for the best results.

Although some of these techniques have produced good estimates of the glottal signal, none of these techniques is completely automated. Glottal inverse filtering can become a useful clinical tool only when the technique requires no user-interaction.

The second indirect method for obtaining information about laryngeal activity is residue inverse filtering. This technique is based on a linear model of speech production, (Atal and Hanauer, 1971; Markel and Gray, 1976; Davis, 1976b). The residue inverse filter is the inverse of the estimated lip radiation, vocal tract and glottal shaping spectral contributions to the speech signal. The result of filtering the speech signal with the residue inverse filter is the residue signal. This signal is an estimate of a periodic source function for an all-pole speech production model, and it exhibits strong peaks at the start of each pitch period and quasirandom noise between the pitch period peaks. Koike and Markel (1975) used the residue signal in a qualitative analysis of normal and pathological voices. They indicated that for some intermediate and advanced cases of laryngeal pathology, the residue signal did not appear to convey more information about the vocal disorder than was already apparent in the speech signal. However, for some early cases, Koike and Markel claimed that the residue signal showed
qualitative evidence of pathology even though the unfiltered speech signal showed no such evidence. Subsequently, Davis (1976b) substantiated these claims by developing quantitative measures based on the residue signal, and verified the hypothesis that more acoustic information about a pathology is conveyed by the (indirect) residue signal than the (direct) speech signal.

**Comparison of Direct and Indirect Signals**

Figure 1 depicts the sound pressure waveform (A), the glottal inverse-filtered signal (B), and the residue inverse-filtered signal (C) for a vowel segment. The inverse filter that produces the residue signal differs from the inverse filter that produces the glottal signal by the addition of several low frequency poles. Thus the glottal signal is equivalent to a low-pass filtered version of the residue signal. In practice, the residue signal is considerably easier to obtain than the glottal signal, since glottal inverse filtering has not been automated. As noted above, the glottal inverse filter coefficients must either be determined from the estimated formant frequencies and bandwidths (Hiki et al., 1976), or the point of glottal closure must be marked for automatic evaluation of the formants (Wong and Markel, 1976). In addition, any low frequency phase distortion will prevent accurate approximation of the glottal signal (especially its closed interval), so the speech signal must be recorded on equipment having a good low frequency phase response even under 100 Hz (Holmes, 1975). A standard audio tape recorder has a poor low frequency phase characteristic, and it is extremely difficult to estimate the glottal signal from a speech signal recorded on such equipment. An FM tape recorder avoids phase problems, but such tape recorders are generally not available in a clinic. In contrast, the residue inverse filter uses a phase-insensitive autocorrelation method to match the overall spectral envelope of the vocal tract, glottal shaping and lip radiation. There is no need for visual inspection of formants, marking of pitch periods, or controlled recording conditions.

In a theoretical sense, the glottal signal has an important advantage over the residue signal, since the glottal signal is a good approximation to the glottal volume velocity waveform (Figure 1), while the residue signal is not directly related to any physically-observable signal. In a practical sense, however, the residue signal may be more easily obtained, and hence has greater potential value than the glottal signal for the clinical assessment of pathological voices.

**Time-Domain Acoustic Parameters**

the sounds was closely correlated with subjective judgments of degree of "roughness."

Measures based on pitch period and amplitude perturbations have been formulated in several different ways. The first of these measures is the "pitch perturbation factor" (Lieberman, 1961, 1963). This parameter is defined as the relative frequency of pitch period perturbations larger than 0.5 ms occurring in a steady vowel sound. A pitch period perturbation is defined as the time difference between the durations of successive pitch periods in the speech signal. Lieberman showed that pathological voices generally have larger perturbation factors than normal voices with comparable fundamental frequencies, and that the perturbation factor is sensitive to the size and location of growths in the larynx.

A second perturbation measure is the "relative average perturbation" (Koike, 1973). This parameter differs from Lieberman's pitch perturbation factor in several respects. Koike observed that steady vowel sounds may normally exhibit slow and relatively smooth changes in pitch period, and he measured rapid perturbations from a smoothed trend line. In addition, Koike suggested normalizing the pitch period perturbation measure by dividing it by the average pitch period. Lastly, Koike suggested that the throat contact signal is a better indicator of laryngeal aperiodicity than the speech signal itself because the effects of the supraglottal structure on the speech sound are reduced. In consideration of these observations, Koike defined the relative average perturbation (RAP) as:

\[
\text{RAP} = \frac{1}{N-2} \sum_{i=2}^{N-1} \frac{P(i-1) + P(i) + P(i+1) - P(i)}{3},
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \frac{P(i)}{P(i)},
\]

where \(P(i), i = 1, 2, \ldots, N\), denote the successive pitch periods. The numerator is the average perturbation measured for each pitch period smoothed by a three-point averaging window, and the denominator is the average pitch period. Koike concluded that the RAP of normal and pathological voices have different ranges, and that the RAP varies significantly between patients with neoplasms and patients with unilateral paralysis.

The basis for a third measure of perturbation is the quasiperiodic amplitude modulation observed in the steady vowel sounds of pathological speakers (Koike, 1969). Koike computed the serial correlation coefficients (correlogram) for the time series of amplitude values at each pitch period peak. He found that the correlograms of normal and pathological speakers are generally distinguishable from one another. The correlograms for speakers with laryngeal tumors usually show significant correlation peaks at lags between three and twelve fundamental periods, and the correlograms for speakers with laryngeal paralysis show no such peaks. Koike concluded that it might be possible to develop methods for differential evaluation of laryngeal pathologies based on information in the amplitude envelope of steady vowel sounds.
In a fourth approach, Kitajima, Tanabe and Isshiki (1975) defined a fundamental frequency (F₀) perturbation measure as the average of 100 successive F₀ perturbations from an all-voiced phrase. An F₀ perturbation is the difference between a measured F₀ and a five-point weighted (in a least-squares sense) average centered around the measured value. A semitone scale (relative to 16.35 Hz or four octaves below middle C) was used, since auditory perception of F₀ is approximately proportional to the logarithm of frequency. They found that normal female voices exhibit larger F₀ perturbation measures than normal male voices, which would be expected since the female F₀ is generally higher than the male. They also found that male voices affected by laryngeal cancer are distinguishable from normal male and female voices by using the F₀ perturbation measure.

Further advances in perturbation measures were made when Takahashi and Koike (1975) combined Koike's earlier results into two time-domain measures of the pathological voice based upon signals obtained from the throat contact microphone; the frequency perturbation quotient (FPQ) and the amplitude perturbation quotient (APQ). The FPQ is defined analogous to the RAP, but the instantaneous F₀ (defined as the reciprocal of the pitch period) is substituted for the pitch period. The APQ is also defined analogous to the RAP, but now peak amplitude values for each pitch period and an eleven-point, rather than a three-point, averaging window are used. Takahashi and Koike found that the APQ made a significant contribution in a principal components analysis of voice quality factors, and that although the FPQ was significantly correlated with the APQ, the FPQ did not make a significant contribution in the principal components analysis.

Davis (1976b) defined a pitch period perturbation quotient (PPQ) and an amplitude perturbation quotient (APQ), which respectively were analogous to Koike's RAP and Takahashi and Koike's APQ. However, there were several differences in the acoustic definitions. Rather than using fixed three-point or eleven-point averaging windows, Davis systematically investigated the benefit of changing the window size, and found that five-point averaging windows for the PPQ and APQ produce the best perturbation measures for normal-pathological discrimination. Davis also found that perturbation measures based on the residue signal are better for discrimination than those based on the speech signal, but he did not attempt any comparisons with the throat contact or glottal signals used by other researchers.

Davis also developed a completely automated procedure for extracting pitch period and amplitude sequences. Once the residue signal is obtained, the extraction procedure uses a peak picking algorithm that finds the significant positive and negative excursions of the signal (Figure 2). The APQs for the positive and negative amplitude sequences are calculated, and the smallest APQ is chosen. The PPQ is then found from the pitch period sequence corresponding to the smallest APQ.

Davis developed two additional time-domain acoustic measures of laryngeal pathology. One measure was based on the observation that the signal-to-noise ratio of the residue signal is a good cue for normal-pathological discrimination (Koike and Markel, 1975; Davis, 1975b). The "signal" in this case is the sequence of spikes spaced at pitch period intervals, and the "noise" is the quasi-random energy between the spikes. Koike and Markel suggested that one
measure of signal-to-noise ratio might be the average of the peak energy for each period divided by the noise energy in the last half of the period for successive pitch periods in a specified interval, but they did not actually attempt any quantitative measurements.

For a residue signal from a normal speaker, the separation of signal from noise for each pitch period is straightforward, and a measure such as the one described by Koike and Markel would suffice. However, for a residue signal from a pathological speaker, the pitch peak is not always distinct from the noise, and it would be very difficult to implement such a measure. The appearance of more noise in pathological cases, and less noise in normal cases, suggested to Davis that the amplitude distribution of the residue signal would be useful for a statistical measure of the signal-to-noise ratio. Figure 3 shows normal and pathological residue signals and the corresponding amplitude distributions. The distribution for the normal speaker is taller and narrower than the distribution for the pathological speaker.

The shape of these distributions may be quantified by a statistical measure called the coefficient of excess (EX) (Cramer, 1958). This coefficient is defined as the ratio of the fourth moment of a distribution to the square of the second moment of the distribution. The EX is zero for a Gaussian distribution. That is,

\[
\text{EX} = \frac{\mathbb{E}\{(x - \bar{x})^4\}}{\mathbb{E}\{(x - \bar{x})^2\}^2} - 3
\]

(2)

where

\[
\mathbb{E}\{(x - \bar{x})^k\} = \frac{1}{N} \sum_{i=0}^{N-1} [x(i) - \bar{x}]^k
\]

(3)

and

\[
\bar{x} = \frac{1}{N} \sum_{i=0}^{N-1} x(i)
\]

(4)

for a signal \(x(i), i = 0, 1, \ldots, N-1\). The EX is positive if the distribution is taller and narrower than a Gaussian distribution and negative if the distribution is shorter and wider. Measurements from numerous speakers substantiate the correlation between the values of the EX and a judgment of the residue signal-to-noise ratio (Davis, 1976b).

The other time-domain acoustic parameter developed by Davis is based on the amount of voicing, or the strength of \(F_0\) during a sustained vowel sound. The detection of \(F_0\) is important in almost any analysis or synthesis study involving speech. In synthesis experiments, for example, the voiced-unvoiced decision is based on the presence or absence of \(F_0\) and is used to determine
whether impulses or noise should be used for the source excitation. One of the oldest digital methods for detecting $F_0$ is autocorrelation analysis (Markel, 1973). Figure 4 shows that a periodic signal, for example, a vowel, exhibits a peak in the autocorrelation function of the residue signal at a duration corresponding to the period. The reciprocal of the period yields the fundamental frequency. Alternately, an aperiodic signal, for example, an unvoiced fricative, shows no pitch period peak. From Figure 4, it is evident that the residue signal is a better indicator of the autocorrelation peak than the speech signal. Davis (1976b) defined a time-domain acoustic parameter called the pitch amplitude (PA) as the value of the pitch period peak in the residue signal autocorrelation function. The PA is high for vowels, small for voiced fricatives, and zero for unvoiced fricatives.

If a given sound is known to be voiced, then the PA becomes a measure of voicing. Voiced sounds from normal speakers have a clearly defined pitch period and the PA is high. In these cases, there is strong periodicity in the glottal volume velocity and area waveforms associated with symmetrical vocal fold movements. Alternatively, "breathy" voiced sounds from pathological speakers are acoustically analogous at the source level to unvoiced sounds from normal speakers. The PA is low or not measurable, the speech sounds have weak periodicity, and there is a significant increase in the amount of noise which is heard.

**Frequency-Domain Acoustic Parameters**

As an alternative to time-domain analysis, frequency-domain analysis provides a different set of acoustic features. A common instrument for frequency analysis of speech is the sound spectrograph, which analyzes the spectral energy distribution of a short speech segment (generally less than three seconds) by filtering the sound with a tracking bandpass filter. The output is a time versus frequency graph of the sound, with spectral energy indicated by intensity. The formants and $F_0$ of a steady vowel are readily visualized in a spectrogram.

Several spectrographic studies show that there are differences between the spectra of pathological voices and the spectra of normal voices (Winckel, 1952, 1954; Nessel, 1960; Yanagihara, 1967a, 1967b; Gould²). The higher frequency harmonics of steady pathological vowels are attenuated in comparison with their normal counterparts. The loss of high frequency harmonics may be caused by changes in the OQ or SQ. In particular, if there is no glottal closure (that is, the OQ is equal to unity), the higher harmonics are sharply attenuated. Spectral noise components may originate in turbulent airflow resulting from incomplete glottal closure or irregular vocal fold vibration (Flanagan, 1958). These spectral noise components are distributed over the spectrum in varying degrees, and the extent of the distribution depends on the severity of the disease. The presence of spectral noise contributes to "hoarseness," which is the first symptom of numerous pathologies; some laryngologists use spectrograms in assessing the degree of and recovery from vocal fold disorders (Rontal, 1975; Gould, 1975). The results of these

---
Figure 1: Comparison of speech, glottal and residue signals for /a/. The glottal signal is closely correlated with the physiological glottal volume velocity waveform (Davis, 1976b).

Figure 2: Extraction of pitch period and amplitude sequences from the residue signal (Davis, 1976b).
Figure 3: Amplitude distribution of normal and pathological residue signals (Davis, 1976b). The decreased noise in the normal residue signal results in an amplitude distribution that is narrow and taller than the pathological amplitude distribution.

Figure 4: Comparison of voiced and unvoiced autocorrelation functions (Davis, 1976b). A pathological voiced sound may be like an unvoiced normal sound, and the PA will be low or nonexistent.
Figure 5: Comparison of long-term average spectra before and after voice therapy (Frejär-Jensen and Prytz, 1976). There is a 3 dB increase in the energy above 1000 Hz relative to the energy below 1000 Hz after therapy.

Figure 6: Voice Profile of subject N1 (Davis, 1976b). The waveforms and measures for this subject exemplify characteristics for a normal healthy voice.
Figure 7: Voice Profile of subject P2 (Davis, 1976b). The subject had unilateral paralysis, and the pathology may have been undetected by auditory perception alone.

Figure 8: Scatter of normal and pathological features, showing two-sigma ellipses for each group (N = normal, P = pathological) (Davis, 1976b).
Figure 9: Comparison of features for voice therapy subject with normal and pathological ellipses (Davis, 1977, see footnote 4). The PPQ, APQ and EX show the best improvement.
studies of sound spectra indicate wide vocal variability, but they also illustrate the feasibility of using the acoustic spectrum to analyze laryngeal pathology.

In an attempt to quantify visual judgments of spectrograms, Hiki et al., (1976) measured the spectral slope of the glottal signal as a feature for acoustic analysis, but they did not present any quantitative results. Other investigators, however, using a "reflectionless" tube (Fisher, Monsen and Engebretson, 1975) indicate that a constant slope is not a particularly good approximation to the glottal spectral envelope even for normal speakers. Hence, the glottal slope of pathological speakers may not be a good parameter for normal-pathological discrimination.

Using a filter bank and a digital computer, Frøkjær-Jensen and Prytz (1976) investigated changes in the long-term average spectrum (LTAS) for patients undergoing therapy for speech disorders such as recurrent nerve paralysis. The LTAS is determined by averaging the spectrum obtained from voiced 80 msec segments of a 45 sec speech sample. They suggested that the ratio of the energy in the 1-5 kHz band to the energy in the 0-1 kHz band (or the decibel difference between the two energy bands) is a good spectral parameter of voice quality. As shown in Figure 5, there is an increase in spectral energy above 1 kHz during therapy, and histograms of the difference between the high and low energy bands (labeled α) indicate approximately a 4 dB increase. Frøkjær-Jensen and Prytz (1976) analyzed more than 50 patients and several normal subjects, and showed that a comparison between high and low energy bands may be used to assess vocal improvement.

Gauffin and Sundberg (1977) found some correlation between LTAS features and perceptual factors such as "overtight-breathy" and "hyper-hypokinetic" obtained in a study by Fritzell, Hammarberg and Wedin (1977). Their LTAS features were decibel energies in the 0-2 kHz, 2-5 kHz and 5-8 kHz bands, and decibel energy differences among the bands. It is noteworthy that Gauffin and Sundberg's energy difference between the 2-5 kHz and the 0-2 kHz bands is nearly identical to Frøkjær-Jensen and Prytz's energy difference, except that the former use 2 kHz as the energy boundary and the later use 1 kHz. The idea of using long-term average spectral measures is a good approach to assessing voice quality and requires additional testing in future investigations.

Davis (1976b) measured normal and pathological spectral characteristics using the concepts of the spectral flatness of the residue inverse filter (SFF) and the spectral flatness of the residue signal (SFR) (Gray and Markel, 1974). Spectral flatness is defined as the ratio in decibels of the geometric mean of the spectrum to the arithmetic mean of the spectrum. Gray and Markel observed that the more noiselike a spectrum, the greater its spectral flatness, having a maximum value of 0 dB for a constant spectrum. Unvoiced sounds, for example, fricatives, which are produced with an open glottis and a significant vocal tract constriction, have greater SFFs than voiced sounds, for example, steady vowels. Since the spectrum of the residue signal is essentially flat, showing only the fine spectral behavior of F_0 and its harmonic components, unvoiced sounds will have lower SFRs than voiced sounds. This result is a consequence of the harmonic nature of normal voiced speech; there are large negative excursions of the residue signal spectrum for each harmonic. As the sound becomes more noiselike (pathological), the harmonic
structure becomes less significant, and the SFR increases.

Using a linear model of speech production, it can be shown that the SFF is the negative sum of the spectral flatnesses of the lip radiation, vocal tract and glottal shaping spectra. If the vocal tract and lip radiation spectra are independent of laryngeal pathology, then changes in the glottal shaping spectrum caused by pathology will be measured by the SFF. It can also be shown that the SFR may be determined by subtracting the SFR from the spectral flatness of the speech spectrum (Gray and Markel, 1974).

Yanagihara (1967a, 1967b) noted the presence of noise components in pathological speech that mask formant characteristics and \( F_0 \) harmonics. Davis (1976b) used this observation as a basis for choosing the SFF and the SFR as spectral measures for vocal assessment. Davis assumed that the SFF is a measure of the masking of formant frequency amplitudes and bandwidths by noise, and that the SFR is a measure of the masking of \( F_0 \) harmonic amplitudes by noise. Since the vocal tract is assumed to be fixed and independent of the source excitation for a steady vowel, these masking effects may be attributed to changes in the sound source harmonic amplitudes caused by variations in the OQ and the amount of source turbulence. The relationship between the OQ and the harmonics of a periodic signal is evident from Fourier analysis; as the OQ increases, the amplitudes of the higher harmonics decrease. Physiologically, the OQ and the amount of turbulence are affected by any pathology interfering with the normal vibratory pattern of the vocal folds. Such effects may be caused by weak muscle action, or changes in the mass, elasticity or compliance of the folds.

VOICE PROFILES

Several studies mentioned above relate acoustic features to listener judgments of voice quality. The results of these studies demonstrate that acoustic parameters such as average pitch period perturbation are significantly correlated with perceptual parameters such as "hoarseness." However, at least one basic question remains. That is, can acoustic parameters alone provide measures that are clinically useful for evaluating pathological conditions in the larynx? For early detection or therapeutic assessment of the pathological voice, it is important to use such parameters in an easily-applied quantitative procedure. Davis (1975, 1976b) suggested that a profile of acoustic characteristics would be as useful to the laryngologist and speech pathologist as an audiogram is to the audiologist. Using features and signals obtained with residue inverse filtering, he developed a Voice Profile to display acoustic information about the voice and to serve as a record in a patient's medical history.

Davis\(^3\) examined the usefulness of the Voice Profile by comparing the visual and numerical information conveyed in the Voice Profiles of normal and pathological subjects (Davis, 1976b) with qualitative observations of the acoustic characteristics of the same subjects (Koike and Markel, 1975). Koike

and Markel described representative residue signals selected from a data base of 10 normal and 10 pathological subjects, and although they indicated that the residue signal could be used to produce acoustic measures of laryngeal function, they did not make any measurements themselves. Davis (1976b) actually made the acoustic measurements and demonstrated that the measures could effectively discriminate between normal and pathological speakers. The data base used by Koike and Markel (1975) and Davis (1976b) is summarized in Table 1, and the acoustic features determined by Davis are listed in Table 2. The Voice Profiles for two of the subjects, N1 and P2, are shown in Figures 6 and 7, respectively. The Voice Profiles display six acoustic features (PPQ, APQ, EX, PA, SFF and SFR) and the signals required to compute the features. The following discussion is based on Davis' 1978 (see footnote 3) comparisons between Koike and Markel's (1975) descriptions and Davis' (1976b) measurements.

Normal Voice Sample

Figure 6 shows the Voice Profile for subject N1. Koike and Markel observed that the speech and residue signals are very regular, with each signal indicating a relatively small amplitude perturbation. They noted that this amplitude perturbation is greater than for subject N7 (not shown), and Davis measured an APQ of 6.08 percent for this subject, as compared to a measured APQ of 2.50 percent for subject N7. Koike and Markel commented that the residue signal has a high signal-to-noise ratio and near-constant periodicity. Davis found that the EX for this subject is 7.69, which is the third highest value in the normal group, and the PPQ for this subject is 0.24 percent, which is the lowest value in both groups. Thus, Davis' acoustic measurements for this normal subject correlate with Koike and Markel's qualitative descriptions.

Pathological Voice Sample

Figure 7 shows the Voice Profile for subject P2. Koike and Markel observed that this slightly hoarse subject represents a case of early laryngeal pathology that possibly would be undetected by auditory perception alone, but that might be detected with good acoustic measures of the residue signal. They indicated that the speech signal shows good periodicity and regularity among pitch periods, but the residue signal shows poor periodicity and a low signal-to-noise ratio. Davis found that the PPQ and APQ for subject P2 are higher than the corresponding measures for all but one of the normal subjects (N5), and the EX is lower than for all but two of the normal subjects (N5 and N10), thus again substantiating Koike and Markel's observations. Also, the observation that the pitch periods are visually similar to one another is confirmed by a high PA of 0.72, a value that is exceeded by only one other pathological subject (P6), and by only three of the normal subjects (N1, N9 and N10).

A comparison of the inverse filter spectra for subjects P2 and N1 shows noticeable differences (Figures 6 and 7). For subject N1, the peaks and valleys of the spectrum are distinct, the bandwidths are small, and the SPF is a low value of -11.3 dB. For subject P2, the valleys between the formant peaks are more shallow, the bandwidths are larger, and the SFF is a higher value of -9.8 dB. These broadband differences would probably be observed in
TABLE 1: Description of normal and pathological subjects age, sex, fundamental frequency and diagnosis.

<table>
<thead>
<tr>
<th>Case</th>
<th>Age</th>
<th>Sex</th>
<th>( F_0 )</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>29</td>
<td>M</td>
<td>110</td>
<td>Normal</td>
</tr>
<tr>
<td>N2</td>
<td>27</td>
<td>M</td>
<td>92</td>
<td>Normal</td>
</tr>
<tr>
<td>N3</td>
<td>30</td>
<td>M</td>
<td>120</td>
<td>Normal</td>
</tr>
<tr>
<td>N4</td>
<td>33</td>
<td>M</td>
<td>138</td>
<td>Normal</td>
</tr>
<tr>
<td>N5</td>
<td>26</td>
<td>M</td>
<td>140</td>
<td>Normal</td>
</tr>
<tr>
<td>N6</td>
<td>24</td>
<td>M</td>
<td>106</td>
<td>Normal</td>
</tr>
<tr>
<td>N7</td>
<td>31</td>
<td>M</td>
<td>124</td>
<td>Normal</td>
</tr>
<tr>
<td>N8</td>
<td>27</td>
<td>M</td>
<td>130</td>
<td>Normal</td>
</tr>
<tr>
<td>N9</td>
<td>23</td>
<td>F</td>
<td>232</td>
<td>Normal</td>
</tr>
<tr>
<td>N10</td>
<td>36</td>
<td>F</td>
<td>180</td>
<td>Normal</td>
</tr>
<tr>
<td>P1</td>
<td>33</td>
<td>F</td>
<td>205</td>
<td>Vocal nodule</td>
</tr>
<tr>
<td>P2</td>
<td>16</td>
<td>M</td>
<td>175</td>
<td>Unilateral paralysis</td>
</tr>
<tr>
<td>P3</td>
<td>56</td>
<td>M</td>
<td>96</td>
<td>Hemilaryngectomized</td>
</tr>
<tr>
<td>P4</td>
<td>25</td>
<td>F</td>
<td>196</td>
<td>Vocal nodule</td>
</tr>
<tr>
<td>P5</td>
<td>39</td>
<td>F</td>
<td>231</td>
<td>Spastic dysphonia</td>
</tr>
<tr>
<td>P6</td>
<td>39</td>
<td>M</td>
<td>115</td>
<td>Vocal polyp</td>
</tr>
<tr>
<td>P7</td>
<td>77</td>
<td>M</td>
<td>164</td>
<td>Laryngeal papilloma</td>
</tr>
<tr>
<td>P8</td>
<td>28</td>
<td>F</td>
<td>189</td>
<td>Unilateral paralysis</td>
</tr>
<tr>
<td>P9</td>
<td>57</td>
<td>M</td>
<td>---</td>
<td>Glottic cancer</td>
</tr>
<tr>
<td>P10</td>
<td>64</td>
<td>M</td>
<td>---</td>
<td>Advanced laryngeal cancer</td>
</tr>
</tbody>
</table>
TABLE 2: Acoustic features for normal and pathological subjects. The six principle features are the pitch period perturbation quotient (PPQ), amplitude perturbation quotient (APQ), pitch amplitude (PA), coefficient of excess (EX), spectral flatness of the inverse filter (SFP) and spectral flatness of the residue signal (SFR).

<table>
<thead>
<tr>
<th>Case</th>
<th>PPQ(%)</th>
<th>APQ(%)</th>
<th>PA</th>
<th>EX</th>
<th>SFP(dB)</th>
<th>SFR(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>0.24</td>
<td>6.08</td>
<td>0.82</td>
<td>7.69</td>
<td>-11.3</td>
<td>-12.3</td>
</tr>
<tr>
<td>N2</td>
<td>0.45</td>
<td>4.17</td>
<td>0.65</td>
<td>8.30</td>
<td>-10.2</td>
<td>-7.6</td>
</tr>
<tr>
<td>N3</td>
<td>0.47</td>
<td>5.42</td>
<td>0.71</td>
<td>3.92</td>
<td>-12.4</td>
<td>-9.4</td>
</tr>
<tr>
<td>N4</td>
<td>0.59</td>
<td>6.70</td>
<td>0.72</td>
<td>2.24</td>
<td>-11.3</td>
<td>-8.7</td>
</tr>
<tr>
<td>N5</td>
<td>0.34</td>
<td>11.63</td>
<td>0.65</td>
<td>1.74</td>
<td>-13.6</td>
<td>-7.8</td>
</tr>
<tr>
<td>N6</td>
<td>0.37</td>
<td>8.72</td>
<td>0.64</td>
<td>5.23</td>
<td>-13.4</td>
<td>-9.4</td>
</tr>
<tr>
<td>N7</td>
<td>0.46</td>
<td>2.50</td>
<td>0.58</td>
<td>11.33</td>
<td>-12.2</td>
<td>-6.3</td>
</tr>
<tr>
<td>N8</td>
<td>0.34</td>
<td>2.56</td>
<td>0.67</td>
<td>7.54</td>
<td>-8.6</td>
<td>-10.4</td>
</tr>
<tr>
<td>N9</td>
<td>0.51</td>
<td>4.59</td>
<td>0.77</td>
<td>3.77</td>
<td>-11.2</td>
<td>-11.9</td>
</tr>
<tr>
<td>N10</td>
<td>5.01*</td>
<td>9.04</td>
<td>0.77</td>
<td>0.96</td>
<td>-11.1</td>
<td>-10.8</td>
</tr>
<tr>
<td>P1</td>
<td>2.61</td>
<td>9.07</td>
<td>0.66</td>
<td>0.63</td>
<td>-9.9</td>
<td>-10.5</td>
</tr>
<tr>
<td>P2</td>
<td>5.08</td>
<td>11.20</td>
<td>0.72</td>
<td>1.90</td>
<td>-9.8</td>
<td>-9.3</td>
</tr>
<tr>
<td>P3</td>
<td>9.60</td>
<td>19.12</td>
<td>0.25</td>
<td>2.05</td>
<td>-8.8</td>
<td>-4.3</td>
</tr>
<tr>
<td>P4</td>
<td>1.87</td>
<td>15.33</td>
<td>0.60</td>
<td>0.97</td>
<td>-6.6</td>
<td>-8.6</td>
</tr>
<tr>
<td>P5</td>
<td>0.85</td>
<td>4.18</td>
<td>0.71</td>
<td>7.27</td>
<td>-5.6</td>
<td>-6.5</td>
</tr>
<tr>
<td>P6</td>
<td>0.60</td>
<td>11.68</td>
<td>0.74</td>
<td>2.98</td>
<td>-8.7</td>
<td>-9.8</td>
</tr>
<tr>
<td>P7</td>
<td>3.29</td>
<td>10.98</td>
<td>0.58</td>
<td>1.07</td>
<td>-13.0</td>
<td>-7.6</td>
</tr>
<tr>
<td>P8</td>
<td>13.25</td>
<td>14.71</td>
<td>0.17</td>
<td>-0.05</td>
<td>-7.4</td>
<td>-5.7</td>
</tr>
<tr>
<td>P9</td>
<td>10.68</td>
<td>16.00</td>
<td>0.49</td>
<td>0.28</td>
<td>-10.9</td>
<td>-7.4</td>
</tr>
<tr>
<td>P10</td>
<td>13.64</td>
<td>15.69</td>
<td>0.26</td>
<td>0.09</td>
<td>-5.3</td>
<td>-3.5</td>
</tr>
</tbody>
</table>

* pitch period tracking errors
visual spectrogram analysis (Yanagihara, 1967a, 1967b; Gould, 1975).

A similar comparison applies between the residue spectra for subjects N1 and P2 (Figures 6 and 7). The harmonic nature of the voice source is readily apparent for the normal subject, and the SFR is a low value of -12.27 dB. In contrast, the residue spectrum for subject P10 exhibits an aperiodic harmonic structure (and hence a more noisy appearance), and the SFR is a higher value of -9.35 dB. These source harmonic differences would also probably be observed in visual spectrogram analysis.

Statistical Analysis of Normal and Pathological Data

In determining the advantages and limitations of these six acoustic features, Davis (1976b) used the data from the ten normal and ten pathological subjects and data from an additional seven normal and eleven pathological subjects in a statistical analysis. The additional subjects had characteristics similar to the first subjects, and were included to increase the population size so that statistical results would be significant. The means, standard deviations and correlations for the normal and pathological groups are listed in Tables 3 and 4.

A t-test (Bruning and Kintz, 1968) shows that the normal and pathological means of the PPQ, APQ and EX are significantly different at the 97.5 percent confidence level, the means of the PA and SFR are significantly different at the 95.0 percent level, and the means of the SFF are significantly different at the 90.0 percent level. Therefore, the PPQ, APQ and EX are the best features for distinguishing between these normal and pathological groups. Also, all of the differences between respective means have the correct sign, for example, the mean normal PPQ is less than the mean pathological PPQ, and the mean normal EX is greater than the mean pathological EX. However, the difference between the normal and pathological means of the SFF is insignificant.

Using Pearson's r score (Bruning and Kintz, 1968), the correlation matrices show two important relationships. For normal and pathological speakers, the correlation between the PPQ and APQ is positive (+0.826) and significant (at the 99.5 percent level). This correlation probably arises from the physical source of abnormal vocal fold vibrations, that is, a change in the mechanical properties of the affected tissues, since this change will cause both pitch period and amplitude perturbations. For normal and pathological speakers, the correlation between the PA and SFR is negative (-0.812) and significant (at the 99.5 percent level). This correlation may be explained as follows. A decrease in the PA indicates more noise and less periodicity in the residue signal (analogous to the generation of unvoiced fricatives), which indicates more noise and less harmonic structure in the residue spectrum, and consequently an increase in the SFR.

In Figure 8, the acoustic features for all subjects are cross-plotted by pairs together with two-sigma ellipses that are derived by a principal components analysis (Davis, 1976b). The axes of each ellipse intersect at the class means and represent orthogonal directions for the scatter of the data. The directions minimize the variance of the data within each normal or pathological class. The nonorthogonal appearance of the axes arises from the
### TABLE 3: Statistics for pooled normal speakers.

<table>
<thead>
<tr>
<th></th>
<th>PPQ(%)</th>
<th>APQ(%)</th>
<th>PA</th>
<th>EX</th>
<th>SFR(dB)</th>
<th>SFF(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.99</td>
<td>7.22</td>
<td>0.725</td>
<td>5.17</td>
<td>-10.50</td>
<td>-11.849</td>
</tr>
<tr>
<td>S.D.</td>
<td>2.28</td>
<td>4.47</td>
<td>0.105</td>
<td>4.29</td>
<td>2.50</td>
<td>1.84</td>
</tr>
</tbody>
</table>

**Correlation**

\[
\begin{array}{cccccc}
& \text{PPQ} & \text{APQ} & \text{PA} & \text{EX} & \text{SFR} & \text{SFF} \\
\text{PPQ} &  & 0.826^* & -0.115 & -0.258 & 0.161 & 0.039 \\
\text{APQ} & -0.100 &  & -0.409 & 0.136 & -0.277 & \\
\text{PA} & -0.015 & -0.812^* &  & 0.164 & \\
\text{EX} &  & 0.262 &  & 0.369 & \\
\text{SFR} &  &  &  &  & 0.018 & \\
\end{array}
\]

* = significant at the 99.5% confidence level

### TABLE 4: Statistics for pooled pathological speakers.

<table>
<thead>
<tr>
<th></th>
<th>PPQ(%)</th>
<th>APQ(%)</th>
<th>PA</th>
<th>EX</th>
<th>SFR(dB)</th>
<th>SFF(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.54</td>
<td>11.99</td>
<td>0.599</td>
<td>2.44</td>
<td>-9.11</td>
<td>-11.851</td>
</tr>
<tr>
<td>S.D.</td>
<td>4.93</td>
<td>6.90</td>
<td>0.236</td>
<td>2.32</td>
<td>3.49</td>
<td>3.35</td>
</tr>
</tbody>
</table>

**Correlation**

\[
\begin{array}{cccccc}
& \text{APQ} & \text{PA} & \text{EX} & \text{SFR} & \text{SFF} \\
\text{PPQ} & 0.625^* & -0.615^* & -0.564^* & -0.491 & 0.064 \\
\text{APQ} & -0.687^* & -0.314 & 0.607^* & 0.070 & \\
\text{PA} & 0.240 & -0.869^* & -0.098 & \\
\text{EX} &  & -0.011 & 0.060 & \\
\text{SFR} &  &  &  & 0.137 & \\
\end{array}
\]

* = significant at the 99.5% confidence level

# = significant at the 99.0% confidence level
use of different scale factors for each axis. In the PPQ-APQ graph, logarithmic scaling is used since linear scaling does not adequately distinguish the normal and pathological classes. The use of logarithmic scaling is noteworthy because Kitajima et al. (1975) suggested that the use of a semitone scale (logarithmically-based) is a better basis for quantifying the auditory perception of F0 perturbation. As a minor point, the mean PPQ and APQ in Figure 8 are different from their respective values in Tables 3 and 4 since the mean of the logarithm of the values is computed for the principal components analysis rather than the logarithm of the mean of the values.

In Figure 8, it is apparent that the normal and pathological classes are best distinguished by the perturbation quotients and least distinguished by the spectral flatness measures. However, even for the perturbation quotients, there are particular points (not indicated) that do not cluster in or near the correct group. Additionally, a normal speaker may have an abnormal value (outside a normal speaker ellipse) in one or possibly two dimensions, while measures in the remaining dimensions may be normal (inside a normal speaker ellipse). For this small number of subjects, it is unrealistic to expect all normal and pathological speakers to fall into tightly-clustered groups. Larger populations grouped by age and sex might yield more representative clusters of normal and pathological data. Also, these results indicate the multidimensional nature of the acoustic detection problem, with some features contributing more information than others for different normal and pathological speakers.

**QUANTITATIVE ASSESSMENT OF VOICE THERAPY**

The usefulness of acoustic features for the assessment of changes in voice quality can be determined by measuring changes in the features over time. In a pilot study, Davis\(^4\) analyzed acoustic features for a patient undergoing voice therapy following removal of a vocal polyp. A trained listener subjectively observed that the voice quality continuously improved during the period of therapy. The data are summarized in Table 5 and compared with the distribution of the earlier data in Figure 9.

<table>
<thead>
<tr>
<th>Date</th>
<th>PPQ(%)</th>
<th>APQ(%)</th>
<th>PA</th>
<th>EX</th>
<th>SFF(dB)</th>
<th>SFR(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/19/73</td>
<td>16.80</td>
<td>10.67</td>
<td>0.86</td>
<td>-0.30</td>
<td>-14.1</td>
<td>-12.7</td>
</tr>
<tr>
<td>6/16/73</td>
<td>4.85</td>
<td>15.51</td>
<td>0.71</td>
<td>1.48</td>
<td>-12.4</td>
<td>-9.6</td>
</tr>
<tr>
<td>8/6/73</td>
<td>0.40</td>
<td>5.97</td>
<td>0.79</td>
<td>6.20</td>
<td>-12.3</td>
<td>-13.0</td>
</tr>
</tbody>
</table>

The PPQ, APQ, and EX are the features in the t-test that show the most significant separation between the normal and pathological mean values. These features also show the best improvement during therapy for this patient. In comparison with the normal ellipses in Figure 9, these parameter values shift from lying outside the normal range at the beginning of therapy to lying within the normal range after several months of therapy. Thus, the results indicate that changes in these features correlate with changes in voice quality for this patient.

None of the other features show changes that correlate with voice quality improvement. For the earliest session, even though the voice quality is poor, and the PPQ is high, the PA is higher than that of any normal or pathological subject listed in Table 2. Visual observation of the speech signal reveals that there is a very high degree of regularity between adjacent pitch periods. The residue signal has a low EX and a very noisy appearance, but the high degree of pitch period regularity observed in the speech signal is maintained in the residue signal and leads to a high PA. Thus the PA may not be as good a measure as the PPQ, APQ or EX as a measure of improvement during therapy.

The SFF and SFR for this subject show no trends that can be correlated with the data used to derive the ellipses. The SFF decreases, and the SFR fluctuates during therapy for this patient. Such trends suggest that these features may be inappropriate for quantifying observations from spectrograms, and perhaps modified or new features would reflect spectral changes more accurately.

Thus, the PA, SFF and SFR show no consistent improvement during therapy and their normal and pathological mean values show less significant separation, but future testing is needed before rejecting them or reducing their weight in an overall assessment of voice quality.

**SUMMARY AND FUTURE INVESTIGATIONS**

The results discussed in this chapter indicate the feasibility of using quantifiable acoustic features to distinguish between normal and pathological subjects. The acoustic features relating to laryngeal function provide more information when inverse filtering is used to remove the supraglottal structure from the speech signal. These features can be automatically computed from a digitized representation of the signal, and the results may be organized to form a Voice Profile.

However, the application of digital analysis techniques poses some difficulties. A linear model of speech production is used to derive the inverse-filtered signal, but the assumption of minimum glottal source-vocal tract interaction may be tenuous for some pathological conditions and requires further study. Nevertheless, the assumption of independence is reasonable for normal and mildly pathological subjects, and these subjects are the ones for whom acoustic analysis is potentially a valuable addition to existing medical procedures.
A further problem is that different sustained vowel sounds will result in different perturbation values (Johnson5). This effect is probably a consequence of the interaction between the glottal source and vocal tract, and indicates the continuing need for study of the complex relationships between the glottal sound source and the acoustic characteristics of the vocal tract. Since any single study uses the same vowel for all speakers, the effect of vowel type on the acoustic measures is probably uniformly distributed among the speakers, and therefore not a significant source of error. Further analysis is also required to determine how consistently acoustic features may be measured from independent samples of the same vowel from the same speaker. Additional statistics should be collected for both sexes and among different age groups.

An additional goal is to analyze acoustic features in a clinic and to detect early cases of laryngeal pathology. Voice Profiles may be useful indicators of voice quality, especially during voice therapy, and it will be necessary to have speech pathologists and physicians evaluate their usefulness. For efficient clinical implementation, tape-recorded voice samples may be sent from a clinic to a central computer facility, either indirectly via the mail system or directly via telephone lines. With the advent of hospital computers and remote terminals in outpatient clinics, an immediate acoustic evaluation of laryngeal pathology is a viable objective. Alternatively, voice samples could be analyzed with a microprocessor-based "black box" built especially for clinical use.

Further study should be directed to at least two problems: the identification of additional features useful in acoustic evaluation of laryngeal pathology, and the possibility of discriminating among types and degrees of pathology. Additional measures might involve the use of long-duration or short-duration voice samples and might include, for example: a) formant frequencies, bandwidths and amplitudes; b) the amplitude of the first harmonic of \( F_1 \) in the residue signal autocorrelation; c) the slope of the line through the peaks in the residue signal autocorrelation; d) the periodicity and peak amplitudes of the pitch period and amplitude correlograms. Some of these features have been suggested previously (Hiki et al., 1976), but neither acoustic or physiological significance nor clinical analysis methods have been established yet for any of these additional features.

Finally, acoustic features should be compared with subjective voice quality ratings (Murry, 1975). Both real and synthesized pathological speech samples (with known acoustic deviations) could be examined. Also, other parameters, for example, the Euclidean distance between a given feature vector and an "ideal" feature vector (having zero perturbation measures, unity pitch amplitude, etc.), might be analyzed as measures of voice quality.

---

It is evident that acoustic voice analysis using inverse filtering may be used for screening individuals for the early detection of laryngeal pathology or for assessing improvement during voice therapy. The fields of speech pathology and laryngology will benefit significantly from future research on the acoustic characteristics of normal and pathological voices.

REFERENCES

Farnsworth, D. W. (1940) High Speed Motion Pictures of Human Vocal Cords. Record 18, Bell Laboratories, 203-208.
Frank, D. I. (1940) Hoarseness - A new classification and a brief report of
four interesting cases. Laryngoscope 50, 472-478.
(Paris: John Wiley and Sons).
Koike, Y. (1969) Vowel amplitude modulations in patients with laryngeal
Markel, J. D. (1971) Formant Trajectory Estimation from a Linear Least-Squares Inverse Filter Formulation, SCRL Monograph no. 7. (Santa Barbara: Speech Communications Research Laboratory, Inc.).
Markel, J. D., A. H. Gray, Jr. and H. Wakita. (1973) Linear Prediction of


Moore, G. P. (1973) *Terminal Report for a Conference on Early Detection of Laryngeal Pathology*. (Gainesville: Communications Sciences Laboratory, Department of Speech, University of Florida).


Tanabe, M., K. Kitajima, W. J. Gould and A. Lambiase. (1975) Analysis of high
speed motion pictures of the vocal folds. Folia Phoniatrica 27, 77-87.
Timcke, R., H. von Leden and P. Moore. (1958) Laryngeal vibrations:
Measurements of the glottic wave. Part I. The normal vibratory cycle.
AMA Archives of Otolaryngology 68, 1-19.
Timcke, R., H. von Leden and P. Moore. (1959) Laryngeal vibrations:
Measurements of the glottic wave. Part II, Physiologic variations. AMA
Archives of Otolaryngology 69, 438-444.
Weiner, N. (1949) Extrapolation, Interpolation and Smoothing of Stationary
Folia Phoniatrica 15, 241-250.
Wendahl, R. W. (1966) Laryngeal analog synthesis of jitter and shimmer,
Winckel, F. (1952) Electroakustische Untersuchungen an der menschlichen
Folia Phoniatrica 5, 232-252.
Wong, D. W. and J. D. Markel. (1976) Considerations in the estimation of
glottal volume velocity waveforms. Journal of the Acoustical Society of
America 59(A), S96-S97.
Yanagihara, N. (1967a) Significance of harmonic changes and noise components
in hoarseness. Journal of Speech and Hearing Research 10, 531-541.
Yanagihara, N. (1967b) Hoarseness: Investigation of the physiological me-
Zemlin, W. R. (1968) Speech and Hearing Science. (Englewood Cliffs:
Prentice Hall).