Workshop on Acoustic Voice Analysis

SUMMARY STATEMENT

BY INGO R. TITZE, PH.D.

NCVS

National Center for Voice and Speech

The National Center for Voice and Speech is a multi-site, interdisciplinary organization dedicated to delivering state-of-the-art voice and speech research to practitioners, trainees and the general public. Members of the consortium are The University of Iowa, The Denver Center for the Performing Arts, The University of Wisconsin-Madison and The University of Utah.

The NCVS gratefully acknowledges its source of support: Grant P60 DC00976 from the National Institutes on Deafness and Other Communication Disorders, a division of the National Institutes of Health.

FORWARD

A workshop was held on the 17th and 18th of February, 1994, in Denver, Colorado to reach better agreement on purpose and methods of acoustic analysis of voice signals. Sponsorship was by the National Center for Voice and Speech, a research and training center funded by the National Institute on Deafness and Other Communication Disorders, and The Denver Center for the Performing Arts. Topics included definitions and nomenclature in voice analysis, algorithms for extraction of parameters, high fidelity recording of microphone signals, computer file structures, sharing of data bases, and development of test signals. Attendance and contributions were by invitation, keeping in mind a balance between industry and academia. The following contributors were present:

David Berry, Ph.D.	University of Iowa and NCVS
Timothy Curran, M.S.	Private Voice Consultant
Dimitar Deliyski, Ph.D.	Kay Elemetrics
Bruce Gerratt, Ph.D.	UCLA VA Hospital
Wolfgang Hess, Dr. - Ing.	University of Bonn, Germany
Yoshiyuki Horii, Ph.D.	University of Colorado and NCVS
David Huang, Ph.D.	University of Washington and Tiger Electronics
Jack Jiang, M.D., Ph.D.	Northwestern University
Issam Kheirallah, M.A.Sc.	University of Western Ontario, and Avaaz Innovations, Inc.
Jody Kreiman, Ph.D.	UCLA VA Hospital
Jon Lemke, Ph.D.	University of Iowa
Martin Milder, B.S.	University of Iowa and NCVS
Paul Milenkovic, Ph.D.	University of Wisconsin, CSpeech, and NCVS
Fred Minifie, Ph.D.	University of Washington and Tiger Electronics
Ed Neuberg, M.S.	Institute for Defense Analysis
Ying Yong Qi, Ph.D.	University of Arizona
David Talkin, B.E.S.	Entropic
Ingo Titze, Ph.D.	University of Iowa and NCVS
William Winholtz, A.A.S.	WJ Gould Voice Research Center,[1] Wintronix and NCVS
Darrell Wong, Ph.D.	WJ Gould Voice Research Center and NCVS

Dr. Wong, Coordinator of Technology Transfer at the National Center for Voice and Speech, acted as chairman of the workshop and editor of the proceedings. Dr. Titze, Director of the National Center for Voice and Speech and Executive Director of the WJ Gould Voice Research Center, led most of the discussions and served as author of the Summary Statement. In this Summary Statement, only the Recommendations (pp 26-30) should be viewed as majority opinion. All other materials are explanatory and the opinion of the author. The full proceedings may be obtained by writing to the National Center for Voice and Speech, Wendell Johnson Speech and Hearing Center, The University of Iowa, Iowa City, Iowa 52242.

[1] *The Wilbur James Gould Voice Research Center is a division of The Denver Center for the Performing Arts.*

CONTENTS

INTRODUCTION

Analysis of acoustic signals of the human voice has many purposes. From a technological standpoint, there is an ever-growing need to store, code, transmit, and synthesize voice signals. The telecommunications industry has dichotomized transmission of information into either *voice* or *data*, suggesting that voice signals are a class of their own. From a basic science standpoint, investigators have traditionally studied the microphone signal to understand speech production and perception, given that the acoustic signal is the common link between them. Finally, from a health science standpoint, the human voice has been shown to carry much information about the general health and well-being of an individual. Our voice reveals who we are and how we feel, giving considerable insight into the structure and function of certain parts of the body.

This workshop was limited to *voice* analysis rather than *speech* analysis, the focus being on the extraction of information about the *source* of sound from a microphone signal. Thus, no attempt was made to discuss or summarize general speech analysis dealing with vocal tract information. For a complete review of speech analysis, the reader is referred to the three volumes of selected papers published by the Acoustical Society of America (Miller et al., 1991; Atal et al., 1991 and Kent et al., 1991).

More specifically, the workshop was a response to an urgency expressed by a group of voice scientists, voice clinicians, and manufacturers of instrumentation to reach some consensus on utility, feasibility, and standardization of voice *perturbation* methods. There has been much expectation and much disappointment in what perturbation analysis can offer for diagnosis and assessment of voice disorders. This workshop gives some of the underlying reasons for both the high expectation and the limited success.

Perturbation analysis is based on the premise that small fluctuations in frequency, amplitude, and waveshape are always present in a voice signal, reflecting the internal "noises" of the human body. Every attempt on the part of the speaker to produce a perfectly steady sound results in an aperiodic waveform. Movements of tissue and air are modulated by the irregular internal motion of electrical impulses, fluids, and cells within an organ. Thus, what might appear to be steady movement or posture on a macroscopic scale is often pulsatile movement on a microscopic scale, as evidenced by twitching of muscles, expansion and contraction of blood vessels, and beating of cilia to transport fluids. If we could shrink to microscopic dimension and travel through the human body, we would see that much of the physical plant (the hydraulic, electrical, and chemical systems) exhibits complex back-and-forth motions (oscillations). These micromovements impose fluctuations on what would otherwise be smooth and steady activity.

Voice production can be thought of as the activation of an entire system of coupled oscillators. The intent to vocalize activates motor commands that are responsible for the neural inputs to an

array of biomechanical, neural, and acoustic oscillators (large box in Figure 1). The vocal folds are the primary oscillating system that produce what we might call the *carrier signal* (the glottal airflow). All other oscillators can then be thought of as *modulators* of the carrier signal. Some of the modulations are nearly sinusoidal (respiratory, heart beat) but many are high dimensional (action potentials of muscles, air vortices, mucus in motion). Yet others are passive oscillators (tracheal resonator, supraglottal vocal tract, various sinuses) that can influence the primary oscillating system.

We can assume that the system of coupled oscillators contains and releases information about the human body; in particular, about its genetics, development, age, disease, language, culture, food and drug intake, and response to the environment (Figure 1). Voice perturbation analysis has the goal of extracting some of this information from the voice signal. The goal is not unlike that of extracting information about the universe from cosmic radiation, or the earth's interior from seismic signals. In all cases, the procedure is extremely difficult and usually requires considerable *a priori* knowledge about the modulations to be extracted.

Therein lies the primary problem of voice perturbation analysis in its present state. We don't know how to measure or classify the multiplicity of perturbations and modulations that are observed simultaneously. Many studies are needed to isolate the individual contributions of each oscillator. Some of these studies are underway (Orlikoff, 1990; Titze, 1991). We also don't know how to apply simple concepts of periodicity and aperiodicity to voice signals. Learning how to quantify aperiodicity is a central focus of this document.

An abundance of terminology tends to mystify what is known about irregularity in voice production. It is appropriate, therefore, to establish working definitions of a few commonly used vocal terms. A more general glossary of terms is included at the end of this summary statement.

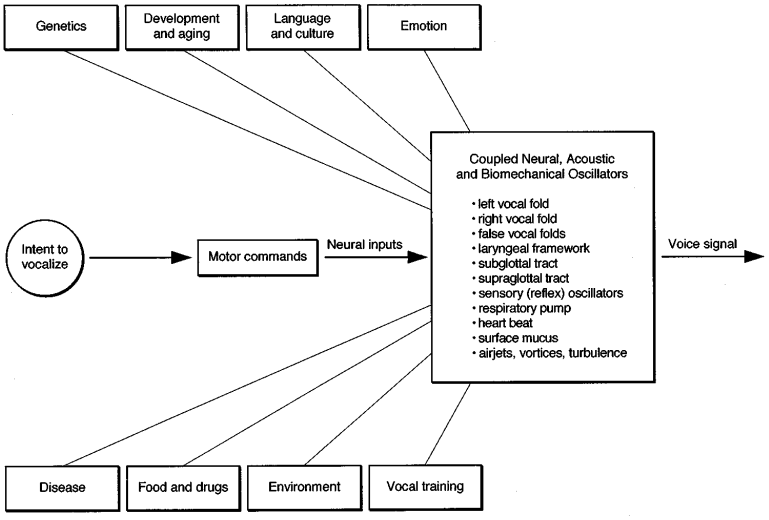


Figure 1.
A list of biological oscillators involved in voice production and factors that may influence them.

NOMENCLATURE AND DEFINITIONS

We begin with a few terms that describe the general phenomenon of irregularity in the human voice, but do not (and probably should not) have precise mathematical definitions.

Descriptive Terminology

A *perturbation* is usually thought to be a minor disturbance, or a temporary change, from an expected behavior. For example, if something is expected to move in a circular orbit but assumes a slightly elliptical path, we say the circular orbit is perturbed. If a person is chewing and encounters a small, hard object in the food, the normal chewing motions are momentarily perturbed. Perturbations are usually such that they do not alter the *qualitative* appearance of a visual or temporal pattern, at least not indefinitely. They are small irregularities that are for the most part overlooked.

A *fluctuation* suggests a more severe deviation from a pattern. It reflects an inherent instability in the system. Whereas a perturbed system usually returns to normal (it is attracted to a stable state), a fluctuating system is somewhat out of control; it cannot find a stable state. Examples are a hand tremor, a flag blowing in the wind, or a car fishtailing on a slippery road. Closer to home in terms of the human voice, a vocal tremor or vibrato may be described as a fluctuation in fundamental frequency and amplitude. It is more than a perturbation because there is no ultimate stabilization of fundamental frequency or intensity toward some constant value. The tremor or vibrato is a pattern itself, rather than a small deviation from a pattern.

Variability is the ability of someone or something to vary, by design or by accident. More formally, it is the amount of variation as determined by a statistical measure. In a golf swing, a basic motion may be repeated over and over again, but conditions of the ground surface, the weather, the ball, the club, or the player may alter the precise motion. Thus, variability may cause the final result (the resting position of the ball) to be far from the expected result. However, depending on how intelligently human variability is used, the final result can also be better than expected. If the player uses variability in muscle activity to compensate for wind and surface variability intelligently, the overall deviation (in the final ball position) may be less than the deviation that would be obtained by a perfectly consistent robot. Thus, variability may be used to fight variability, but it can also have a catastrophic effect if allowed to run rampant. (For a discussion of variability in speech, see Perkell & Klatt, 1986).

Jitter refers to a short-term (cycle-to-cycle) perturbation in the fundamental frequency of the voice. Some of the early investigators (e.g., Lieberman, 1961, 1963) displayed speech waveforms oscillographically and saw that no two periods were exactly alike. The fundamental frequency appeared jittery; hence, the term jitter. *Shimmer* was then invented as a companion word for amplitude-jitter; i.e., a short-term (cycle-to-cycle) perturbation in amplitude (Wendahl, 1966).

A problem has arisen in trying to make a precise mathematical definition stick for jitter or shimmer. What is meant by short term, for example, and what kind of variability measure should be adopted? There are many ways of quantifying a deviation from an expected pattern or trend. This has led to a proliferation of mathematical definitions for jitter and shimmer. We believe that it is best to leave the terms as they are (as generic descriptors of fundamental frequency and amplitude variability) and use more standard terminology of engineering and statistics to quantify error measurements (see the later section on perturbation measures).

An unfortunate misunderstanding can arise for singing teachers who use the term shimmer to describe a beautiful bell-like vocal quality. A shimmering voice is aesthetically most pleasing in this context. As a random short-term amplitude perturbation, however, shimmer is not particularly pleasing to listen to. It is usually perceived as a crackling or buzzing sound, and in extreme cases, it can become very unpleasant and rough. It is important to communicate, therefore, the context in which the term shimmer is used.

Tremor is a low-frequency fluctuation in amplitude or frequency (or both). Its origin is usually neurologic. Physiologic tremors in the body have fluctuation rates between 0-15 Hz, but not all are perceived the same way auditorily when they are part of the vocal signal. Thus, a low-frequency tremor (0-3 Hz) is perceived as a *wow*. This is also the term used by the recording industry to describe variability in the speed of the tape drive of an audio recorder. A companion term, *flutter*, describes the variability associated with tape contact on the recording head. In the voice literature, flutter has been used to describe neurologic fluctuations in the 9-15 Hz range (Aronson et al., 1992). Flutter appears to be associated with rapid onset and offset of phonation, reflecting the natural oscillating frequency of the adductor-abductor control system in phonation. Some singers tend to cultivate this natural frequency in the production of *trillo* - a fast, fluttering ornament typically used in renaissance music (Hakes et al., 1990).

In the mid-range rate (4-8 Hz), vocal tremor is part of the natural quality of the human voice, provided it's extent does not exceed certain limits. Synthesis has shown that without a small degree of tremor, steady vowel production has a buzzy quality. There is something about a low frequency fluctuation in the voice that makes it warm and acceptable. An exaggerated extent of vocal tremor, on the other hand, is considered pathologic (Koda & Ludlow, 1992).

The origin of vocal *vibrato* is not completely understood, but some evidence is beginning to show that vocal vibrato may be a stabilized physiologic tremor in the laryngeal muscles (Niimi et al., 1988; Ramig & Shipp, 1987). It is conceivable, though speculative at this point, that a natural vocal vibrato can be cultivated from a 4 to 6 Hz physiologic tremor in the cricothyroid and thyroarytenoid muscles. This would require some mechanical load or reflex loop to stabilize irregular movement (Titze et al., 1994).

For the description of pathologic voices, several terms have found universal appeal. *Roughness* refers to an uneven, bumpy quality. It results from irregularity in the energy contained in a critical band of the auditory system (Terhardt, 1974). Periodic sounds (such as vocal fry) can have

roughness, but more often there is a lack of periodicity. *Breathiness* is a vocal quality that contains the sound of breathing (expiration, in particular) during phonation. Acoustically, there is a significant component of noise in the signal due to glottal air turbulence. Sometimes the term *hoarseness* is used to describe the combination of roughness and breathiness.

The terms described thus far - perturbation, fluctuation, variability, jitter, shimmer, tremor, wow, vibrato, flutter, roughness, breathiness, hoarseness, and several others defined in the glossary - have no mathematical definitions. No numbers or physical units of measurement need to be attached to them, although some of them can be rated psychophysically. Nevertheless, they serve a purpose in describing vocal phenomena and the associated physical processes. At this point, some additional terms will be reviewed that have mathematical definitions.

Periodicity, Subharmonics, and Modulation

A series of events is termed periodic if the events cannot be distinguished from one another by shifting time forward or backward by a specific interval nT_o,

$$f(t \pm nT_o) = f(t) \qquad (1)$$

where n is any positive integer and T_o is the *period*. T_o must be the smallest value possible to be deemed the *fundamental period*. Equation (1) can never be strictly satisfied in a voice signal. All vocal events tend to be aperiodic. The term *quasi-periodic* is sometimes used to suggest that there is only a small deviation from periodicity. It must be kept in mind, however, that quasi-periodicity is simply a special case of aperiodicity. Furthermore, in physics the term quasiperiodic has the special meaning of the superposition of two or more periodic signals with incommensurate (non-integer ratio) frequencies. Hence, we prefer not to use the term, but adopt *nearly-periodic* to avoid confusion.

A series of events is termed *cyclic* if the events recur, but not necessarily in periodic fashion. A cyclic event is recognized on the basis of a pattern that involves neighboring points on a waveform (e.g., a zero crossing, a maximum value, a minimum value).

A *cyclic parameter* is a construct of cyclic events (e.g., inter-pulse-interval, open quotient, skewing quotient, peak-to-zero amplitude, peak-to-peak amplitude, maximum flow declination rate). Some of these parameters are identifiable only after the acoustic waveform has been *inverse filtered*, which is the process of removing the vocal tract resonances from the waveform to obtain the glottal airflow (Rothenberg, 1973). In a sinusoidal waveform, the amplitude A, the period T, and the frequency 1/T are obvious cyclic parameters and have precise definitions. In a complex periodic waveform, the fundamental period T_o and fundamental frequency $F_o = 1/T_o$ also have exact definitions (equation 1), but amplitude can be defined in a variety of ways. Traditionally, the peak value (maximum positive or negative) and the peak-to-peak value (maximum positive to maximum negative) have been used. As alternatives, Hillenbrand (1987) used the root-mean-squared (RMS) intensity in

each cycle as the representative amplitude, while Milenkovic (1987) used a gain factor *k* calculated as part of a cycle to cycle least squared error comparison.

A *cyclic parameter contour* is a time series of any cyclic parameter (e.g., F_o contour, amplitude contour, open quotient contour). For periodic signals, the contour is a constant, by definition. For aperiodic signals, the cyclic parameter contour can take on many different shapes, becoming a signal of its own. Figure 2 shows an F_o contour extracted from a voice signal (top curve). The F_o contour is highly magnified to show the finest detail in perturbation. The subject (normal male) sustained an [b] vowel as steady as he could for about 12 seconds at a mid-range value of 99.8 Hz. The target F_o was 98 Hz, a G_2 on the keyboard. Time is labeled in number of cycles (1195 total) instead of seconds because 1 point is plotted for every cycle of vocal fold vibration. Note that the range of frequency variation is 96.7 Hz to 102.4 Hz, about ±3%, but this range is attributed mainly to

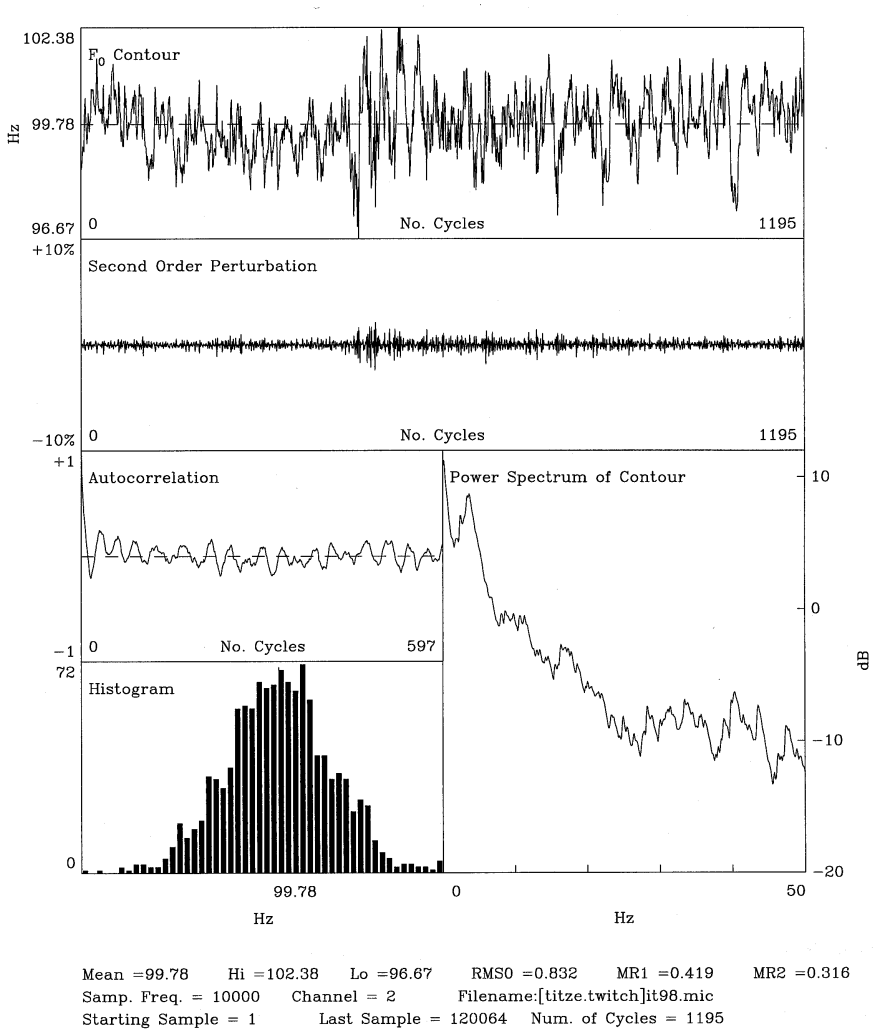


Figure 2. A fundamental frequency (F_o) profile used for perturbation analysis. The subject was a normal adult male phonating a steady [b] vowel at approximately 100 Hz for about 12 seconds.

one burst of instability in the middle of the contour. Over the rest of the utterance, the F_o variation was considerably smaller. (Other graphs in Figure 2 will be discussed later).

Now let x_i represent an arbitrary cyclic parameter, for which some stylistic contours are illustrated in Figure 3. Part (a) shows an irregular contour, similar to that of Figure 2 just discussed, but with fewer cycles. Part (b) shows a regular "up-down" pattern that is often seen in voice signals, and parts (c) and (d) show a linear and sinusoidal trend, respectively. The "up-down" pattern in part (b) suggests the presence of a *subharmonic frequency* $F_o/2$, or a *period doubling* $2T_o$. Clearly, if only every other point were plotted in the contour, a constant would result and periodicity would be achieved. Thus, the true period is doubled. In equation (1), period doubling is represented by using only the even values of *n*.

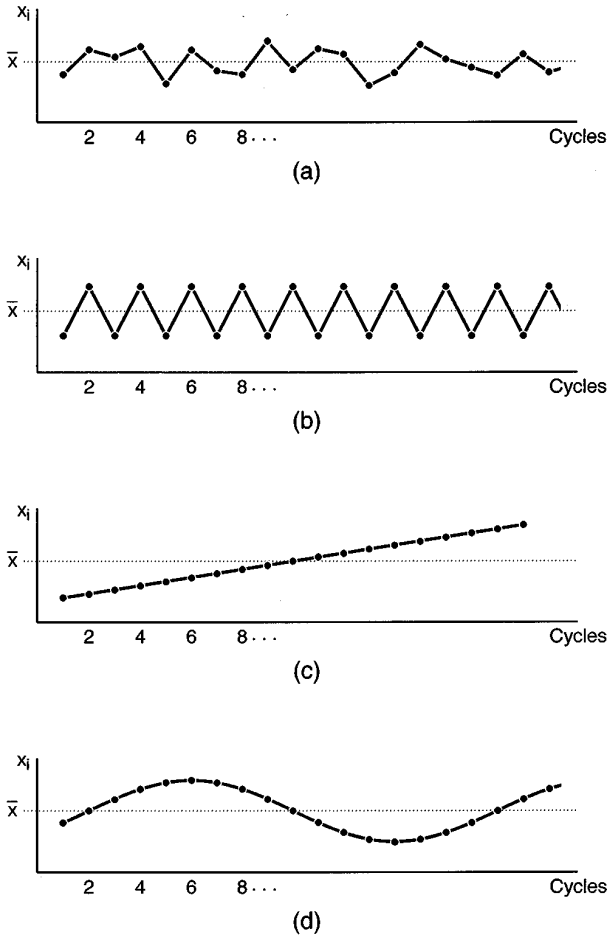


Figure 3. Modulations of a cyclic parameter x_i around the mean value (a) random, (b) alternating, (c) linear trend, and (d) sinusoidal.

The "up-down" sequence is also referred to as a *period-2 sequence* in nonlinear mechanics. This nomenclature can be extended to define a *period-3 sequence* (the pattern would be high-low-middle) or to a *period-4 sequence* (high-low-very high-very low), and so on. The terms *diplophonia, triplophonia, quadruplophonia* have also been used in the description of these sequences, but the terminology has not been universally adopted. In general, a *period-n sequence* in the parameter contour would be called *multiplophonia* if it were important to retain reference to the word "phonation" in the nomenclature. However, "period-n phonation" or "phonation with an F_o/n subharmonic" accomplishes the same objective.

But why isn't F_o/n simply redefined as the fundamental frequency? That depends on the relative energy contained in the subharmonic. Often the period-n variations of a cyclic parameter are small, suggesting that "on average" the cyclic parameter has not changed. Furthermore, the auditory perception of the cyclic parameter (e.g., pitch in the case of F_o or loudness in the case of amplitude) may not have changed, but rather a dimension of roughness or some other quality has been added. Their frequencies are commensurate (in integer ratio) with the primary frequencies and may or may not be perceived as separate pitches.

In contrast to period-n phonation or multiplophonia, the term *multiphonia* is used to suggest the presence of several independent phonations (sound sources). Thus, *biphonia* would contain two independent sources, such as the true vocal folds and the false vocal folds, and *triphonia* would contain three independent sound sources (perhaps the addition of a glottal whistle). Their frequencies would not have to be commensurate. However, different modes within the same sound source may also generate independent frequencies, making the identification of the sound sources a nontrivial matter.

The term *modulation* is used to quantify the systematic change of a cyclic parameter (usually frequency or amplitude) of a periodic signal. The periodic signal (usually a sinusoid) is called the *carrier* in communication theory. In phonation, the carrier is the sequence of periodic airflow pulses emitted from the glottis, and the modulation is the slower variation of cyclic parameters discussed in the previous section. In radio communication, the entire voice signal modulates an electronically generated sinusoid for wireless transmission (typically in the MHz range), suggesting that modulations can be stacked up (layered) upon each other. The carrier of one signal becomes the modulation of another.

Figure 4a demonstrates an amplitude modulation (*AM*) and Figure 4b a frequency modulation (*FM*) of a series of glottal pulses. Mathematically, the *modulation extent* is defined as

$$E_{AM} = \frac{A_1 - A_2}{A_1 + A_2} \tag{2}$$

for sinusoidal amplitude modulation, and

$$E_{FM} = \frac{T_1 - T_2}{T_1 + T_2} \tag{3}$$

for sinusoidal frequency modulation, where A_1 and A_2 are the largest and smallest amplitudes, respectively, and T_1 and T_2 are the largest and smallest periods in the signal. Note that modulation extent approaches 1.0 (100%) when either A_2 or T_2 approaches zero. Such an extreme condition violates a basic principle of modulation, however, because the carrier signal momentarily loses its amplitude completely for *AM*, whereas the frequency (1/*T*) momentarily approaches infinity for *FM*. Practical modulations are usually well below 100%. In a vocal vibrato, for example, a 3% frequency modulation is typical. Amplitude modulations can be larger in vocal signals, but seldom exceed 50%.

For modulation extent to be measurable in a voice signal, the *modulation frequency* F_M (the number of modulation cycles per second) should be well below the carrier frequency $F_c = F_o$. (In the theoretical limit, F_m/F_c is governed by the Nyquist frequency). If F_m is too high, there is insufficient sampling of the modulation envelope and large errors may occur in its detection. Such is the case with subharmonic modulations, which are often undersampled in a voice signal (note that there are only two points per cycle in Figure 3c). Vibrato and tremor, on the other hand, are usually adequately sampled because their frequencies are naturally well below F_o (see Figure 3d as an example).

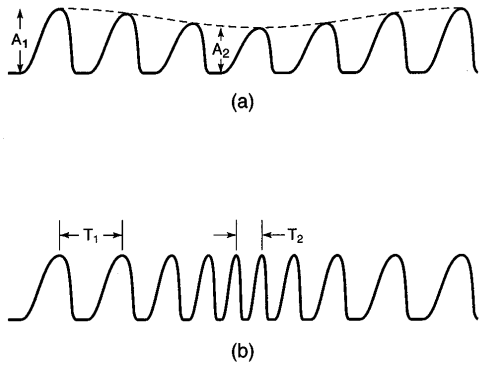


Figure 4. (a) Amplitude modulation (AM), (b) frequency modulation (FM) of a series of glottal pulses.

Perturbation Functions

As before, let x_i to be a cyclic variable of vocal fold vibration that has been extracted from the i-th vibratory cycle. A *window* of observation is defined, containing *N* cycles of vibration, so that the subscript *i* can range from 1 to *N* in the observation window.

The *mean value* of the cyclic variable over the window of observation is defined as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad . \tag{4}$$

If the mean value is intended to be a constant, as in steady vowel phonation, then a *zeroth-order perturbation* of the i-th cycle can be defined as

$$P_{0i} = x_i - \bar{x} \quad . \tag{5}$$

(The term zeroth-order is used because a constant is basically a zero-order pattern or trend). Higher-order perturbation functions are defined as the following finite differences:

$$P_{1i} = x_i - x_{i-1} \tag{6}$$

$$P_{2i} = x_i - \frac{1}{2}(x_{i+1} + x_{i-1}) \tag{7}$$

$$P_{3i} = x_i - \frac{1}{3}(x_{i+1} + 3x_{i-1} - x_{i-2}) \tag{8}$$

$$P_{4i} = x_i + \frac{1}{6}(x_{i+2} - 4x_{i+1} - 4x_{i-1} + x_{i-2}) \tag{9}$$

In general, since the first subscript represents the order *n* of the perturbation function and the second subscript represents the i-th cycle, higher-order (n+1) perturbation functions are generated recursively as

$$P_{n+1,i} = \frac{1}{K}\left(P_{n,i} - P_{n,i-1}\right) \qquad n = 0, 2, 4... \tag{10}$$

$$= \frac{1}{K}\left(P_{n,i+1} - P_{n,i}\right) \qquad n = 1, 3, 5..., \tag{11}$$

where K is a normalization factor that keeps the coefficient of x_i positive and unity in each perturbation function. Note that with this normalization, all perturbation functions are zero when x_i is a constant.

The perturbation functions can be used to remove known or assumed trends in the cyclic parameter contour. The zeroth-order perturbation function removes nothing, the first order perturbation function removes a constant (the mean value $\bar{x}$), the second order function removes a linear trend, the third order function removes a quadratic trend, and so on. In general, the n-th order perturbation function removes a polynomial trend of order n-1 in the contour.

Consider a linear trend as shown in Figure 3c. It is represented by the relation

$$x_i = x_{i-1} + k \quad , \tag{12}$$

where k is the rise per cycle. It is easily seen from equation (6) that $P_{1i} = k$ and that all higher-order perturbation functions in this example are zero. Thus, the first order perturbation function *extracts* the linear trend, whereas the higher order perturbation functions *remove* it. The second graph from the top in Figure 2 shows a second order perturbation function computed from a human voice. The scaling is smaller than that of the contour because it is an absolute scaling ($\pm$10% deviation from the mean value). Note that the short-term fluctuations of the contour are retained, but the long-term trends are removed. For example, the gradual downward slope of the F_o contour in the beginning one-third of the utterance has been removed. So has the tremorous variation that is most noticeable in the middle of the contour. All that is left in the second-order perturbation is the short-term "noise".

If a linear trend is deliberately produced by the voice, such as a uniform F_o glide between two pitches in a specified amount of time, then k is a known quantity. It can simply be inserted into the perturbation formulas. For example, the first-order perturbation then becomes

$$P_{1i} = (x_i - x_{i-1}) - k \quad , \tag{13}$$

which is now known as the *deviation from a linear trend.* If a linear trend is suspected as an inherent pattern, but k is not known, it can be computed from the data by linear regression. This is a well-known statistical procedure (Hays, 1988). Furthermore, all patterns with forward predictability (e.g., a sinusoid, a damped sinusoid, an exponential) can collectively be removed by linear predictive coding (LPC), with only random (or unpredictable) events remaining in the residual perturbation function. LPC analysis is based on the assumption that x_i can be predicted from a weighted sum of M previous samples,

$$x_i = \sum_{j=1}^{M} a_i x_{i-j} \quad , \tag{14}$$

where the a's (the predictor coefficients) are determined by a linear least squares fit to the contour (Markel & Gray, 1976).

Some investigators have opted to use a hybrid between the zeroth-order perturbation function and the second-order perturbation function,

$$P_i = x_i - \frac{1}{2m+1} \sum_{j=-m}^{m} x_{i+j} \qquad (15)$$

This function computes the *deviation from a local mean*. If 2m + 1 = N, the total number of cycles in the window, the perturbation function becomes P_{oi} (equation 5). If m = 1, then the summation becomes the three-cycle local average used by Koike (1973). For a two-cycle local average, the j = 0 value is omitted and P_{2i} is obtained (equation 7). An 11 cycle average (m = 5) has also been used (Takahashi & Koike, 1975).

The *autocorrelation function* of the cyclic parameter contour serves a purpose contrary to that of a trend remover (such as the second-order perturbation function). It removes the short-term cycle-to-cycle "noise" but keeps the long term patterns. Mathematically, the autocorrelation function is computed as

$$c_i = \sum_{j=1}^{N/2} x_j x_{i+j} \Big/ \sum_{i=1}^{N/2} x_i^2 \qquad i = 0, N/2 \quad . \qquad (16)$$

where the brackets indicate average (expected) values over a fixed window of observation. Basically, the autocorrelation function is the contour multiplied by a delayed version of itself, the delay being one period, two periods, three periods, and so on (Rabiner & Schafer, 1978; Bendat & Piersol, 1986). In Figure 2 (third waveform from top on left side), the computation was done from 0 delay periods to 597 delay periods. The autocorrelation is always maximum for 0 delay periods (the function correlates perfectly with itself if not delayed), where it has the value 1.0. At all other points, it is greater than -1.0 and less than +1.0 if properly normalized. Note that the fluctuation seen in the autocorrelation function indicates that a small amount of a "vibrato" is present in the subject's voice. This is perceptually below threshold. The subject intended to produce a straight tone, but since he was vocally trained to sing with vibrato, he could not completely suppress it. This is a good example, then, of a case in which acoustic analysis "digs out" something that is easily lost in both the raw F_o contour and the auditory perception.

The *histogram* (bottom left corner in Figure 2) shows a distribution of the cyclic parameter values for all of the 1,195 cycles. On the vertical axis is the number of the occurrences of the parameter value in a narrow range (bin). Note that the greatest number of occurrences of F_o are near the midrange value (99.8 Hz), whereas large deviations from the midrange occur infrequently. The distribution is nearly Gaussian, suggesting that perturbations are primarily random. In contrast, the distribution would be bimodal (two major peaks) if a subharmonic or a strong vibrato were present in the F_o contour.

Finally, the *power spectrum of the parameter contour* (bottom right) is a useful display of the dominant frequencies that modulate the contour. Note that a frequency of about 5 Hz stands out in this spectrum. This is the frequency of the small amount of vibrato in the voice. All other peaks in the power spectrum are at least 10 dB lower and do not represent significant components. Again, subharmonics, tremors, or any other modulations can easily be detected in this type of display.

In summary, a cyclic parameter profile of the type shown in Figure 2 is a useful tool in voice analysis. It helps to quantify visually what is perceived aurally. A similar profile can be constructed for amplitude variation or for any other cyclic parameter (open quotient, maximum flow declination, skewing quotient, etc.).

Perturbation Measures

A perturbation measure is an effective value of the overall perturbation in the cyclic contour. For example, the standard deviation from the mean is

$$\sigma_o = \left[\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2\right]^{1/2} \quad . \tag{17}$$

This measure can also be identified as the *root-mean-squared* (RMS) value of the zeroth-order perturbation function (recall equation 5) .

The *mean rectified value,* or *mean absolute value,* of the zeroth-order perturbation is defined as

$$\delta_o = \frac{1}{N}\sum_{i=1}^{N} |x_i - \bar{x}| \quad . \tag{18}$$

This measure of perturbation is fundamentally not much different from σ_o, but it is a little easier to compute because it does not involve squares and square roots. Also, it does not weight outliers (large deviations from the mean) as heavily as σ_o because first-power terms rather than second-power terms are used in the summation.

In general, a collection of perturbation measures can be written as

$$\sigma_n = \left[\frac{1}{N-n}\sum_{i} P_{ni}^2\right]^{1/2} \tag{19}$$

$$\delta_n = \frac{1}{N-n}\sum_{i} |P_{ni}| \quad , \tag{20}$$

with δ_1 being the most frequently used measure in the literature. In Figure 2, σ_o has the value of 0.832%, σ_1 has the value of 0.419%, and δ_2 has the value of 0.316%.

Both δ_n and σ_n are *magnitude* perturbation measures only. The squaring and absolute magnitude operations remove all information about the *direction* in which the cyclic variable deviates from the mean value. Consider again the four contours shown in Figure 3. They appear quite different visually but could all produce rather similar perturbation measures. The magnitude perturbation measures σ_n and δ_n tell us little about the *patterns* in the perturbations functions. They are totally insensitive to any regularity that may exist in the deviations. Indeed, the only pattern they relate to is a constant, the mean value $\bar{x}$. This is a serious limitation for many applications in voice perturbation analysis because the patterns may reveal more about the nature of a disorder, or special voice characteristic, than a simple magnitude error measure. (For a more detailed discussion of magnitude versus directional perturbation measures, see Pinto & Titze, 1990).

Several investigators have used a harmonics to noise ratio (Yumoto et al. 1982; Cox, 1989), a signal to noise ratio (Klingholz, 1987), or a normalized noise energy (Kasuya et al. 1986) to quantify the aperiodic portion of the voice signal. The harmonic energy is first defined as

$$E_h = N \int_0^T f_A^2(\tau)\, d\tau \quad , \tag{21}$$

where N is the number of cycles, T is the greatest period found among the N cycles, and f_A is the average acoustic waveform per cycle (obtained by padding all cycles to the maximum period with zeros and averaging point by point from event marker to event marker). The noise energy is then defined as

$$E_n = \sum_{i=1}^{N} \int_0^T \left[f_i(\tau) - f_A(\tau) \right]^2 d\tau \quad , \tag{22}$$

where f_i is the waveform in the i-th cycle, and the harmonics to noise ratio is

$$HNR = 10 \log_{10}(E_h / E_n) \quad . \tag{23}$$

If the HNR is used as a perturbation measure, it needs to be noted that this measure is not specific to any cyclic parameter. Therein lies its asset as well as its liability. One cannot tell if the period, the amplitude, or the waveshape is perturbed. Simple Gaussian noise added to a periodic waveform can decrease the HNR, as will jitter or shimmer. Thus, the measure correlates best with an overall perception of "noisiness and roughness" in the signal, regardless of what the source might be. New approaches described by Qi (1992) and Qi et al. (1995) includes a time-base correction that minimizes the effect of jitter as a contributor to noise. Thus, these approaches begin to separate the sources of noise in the HNR measure.

SIGNAL TYPING

The most interesting voice signals are encountered when vocal fold vibration is highly influenced by nonlinearity in tissue and air movement, or when coupled oscillator modes become desynchronized. For example, two modes of the same vocal fold, or two modes between opposite folds, may compete for dominance. A resolution to the mode conflict is what we have described as period-n phonation, whereby each mode is allowed to have its turn, so to speak, making the overall period much longer. Another resolution is a long-range modulation (over several cycles), the frequency of which is incommensurate with F_o. In some cases, however, there is no resolution at all in terms of any real or apparent periodicity, and oscillation becomes chaotic.

In the language of nonlinear dynamics, a qualitative change in the behavior of a dynamical system is known as a *bifurcation*. It usually occurs when some parameter of the vibrating system is changed gradually (e.g., lung pressure, vocal fold tension, or asymmetry between the vocal folds). Figure 5 shows sketches of how glottal flow waveforms transform after two successive bifurcations. The first bifurcation is seen as a period doubling (part *a* to part *b*) whereas the second is seen as a total loss of periodicity (part *b* to part *c*).

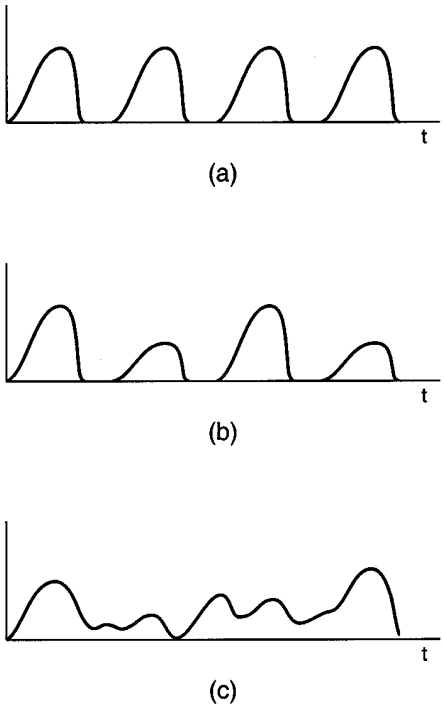


Figure 5. A series of glottal pulses showing evidence of bifurcation. (a) periodic vibration, (b) period doubling, (c) chaotic vibration.

The following classification scheme is adopted here to recognize the nature of bifurcations in voice signals. The classification is central to all other considerations in acoustic voice analysis. It follows the general principles of nonlinear dynamics of coupled oscillators.

Type 1 signals - nearly-periodic signals that display no qualitative changes in the analysis segment; if modulating frequencies or subharmonics are present, their energies are an order of magnitude below the energy of the fundamental frequency.

Type 2 signals - signals with qualitative changes (bifurcations) in the analysis segment, or signals with subharmonic frequencies or modulating frequencies whose energies approach the energy of the fundamental frequency; there is therefore no obvious single fundamental frequency throughout the segment.

Type 3 signals - signals with no apparent periodic structure.

A spectrogram is useful in making the classification. For example, Figure 6 shows a spectrogram of a patient with hyperfunctional childhood dysphonia. The fundamental frequency is 300 Hz. Bifurcations can be seen to occur around 400 ms (the beginning of a period-3 phonation), around 900 ms (return to the original), and around 1100-1200 ms (beginning of a mixture between period-3 and period-4 phonation). The signal is therefore classified as type 2.

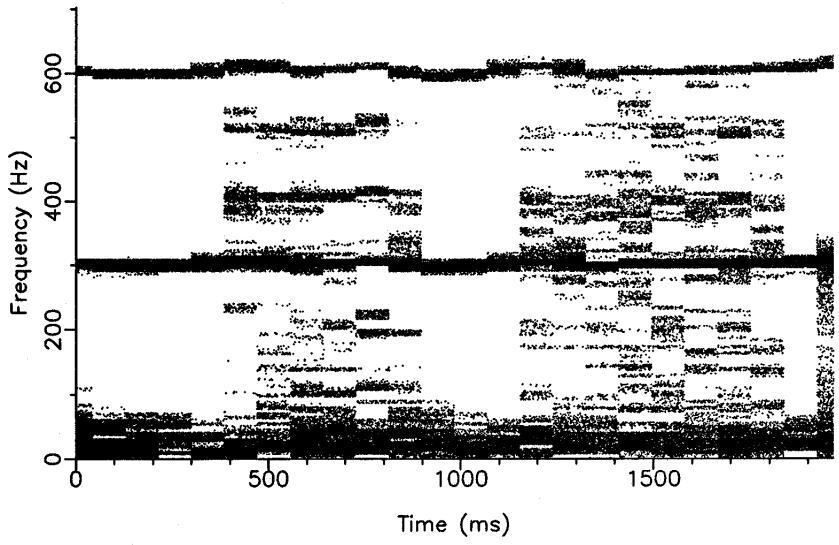


Figure 6. Narrow-band computer spectrogram for a patient with hyperfunctional childhood dysphonia. Abrupt transitions to different phonatory regimes are visible, indicating bifurcations in vocal fold vibration.

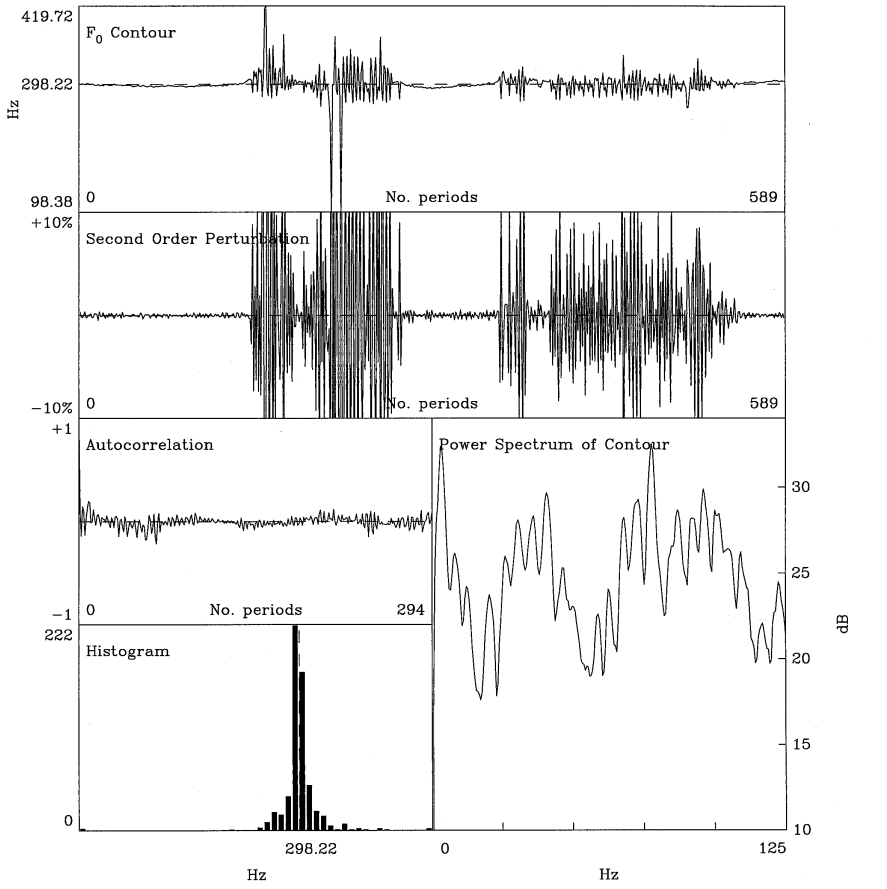


Figure 7. Fundamental frequency (F_o) profile for the patient with hyperfunctional childhood dysphonia.

A fundamental frequency profile, similar to that of Figure 2, is shown for this dysphonic patient in Figure 7. Note that bifurcations can be identified in the F_o contour as segments where the F_o extractor is uncertain about the constant 298 Hz value. In two cycles the extracted F_o drops down to 98 Hz, close to the $F_o/3$ subharmonic. In one case, the extracted F_o jumps to 420 Hz. In general, F_o is extracted reliably only in the three segments where the waveform is nearly periodic.

The second-order perturbation function has wild fluctuations. It is clear from this display that a single perturbation measure for the entire segment is meaningless and that the visual displays carry more information than can be characterized by a single number.

As another example, analysis was performed on the waveform of a patient with unilateral laryngeal nerve paralysis (Figure 8). The waveform itself shows intermittent segments of low frequency modulation (segments b and d). The fundamental frequency is 285 Hz and the modulation frequency is 32 Hz. If only segments a, c, and d had been acquired and analyzed, the signal would have been classified type 1. As it is, it is clearly a type 2 signal.

Figure 9 shows its corresponding narrow-band spectrogram. The 32 Hz modulation is seen as closely-spaced horizontal lines on both sides of the three harmonics, i.e., as sideband frequencies. These frequencies are not in exact integer ratios of F_o. There are between 8 to 10 short lines between each of the long lines.

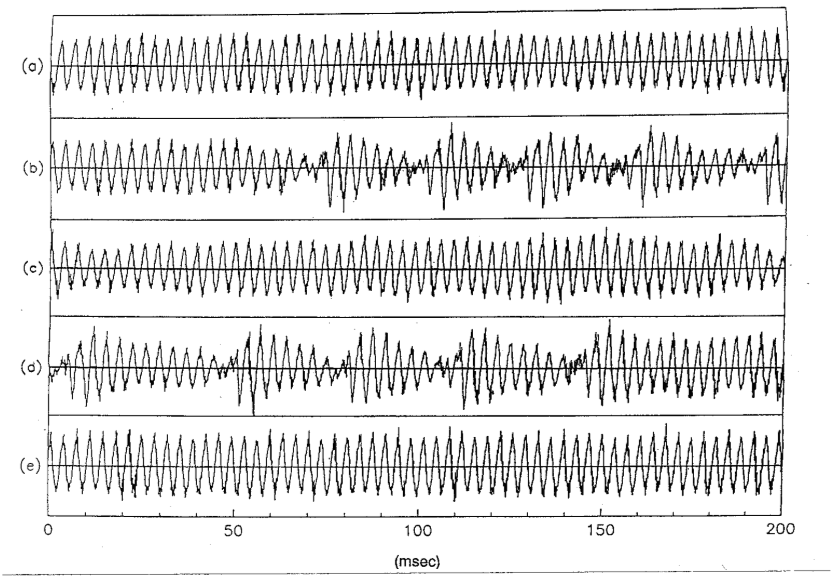


Figure 8. Microphone signal of a patient with unilateral lyarngeal nerve paralysis. Parts (a) to (e) should be viewed serially, 200 ms per segment, for a total of 1s (After Herzel et al, 1994).

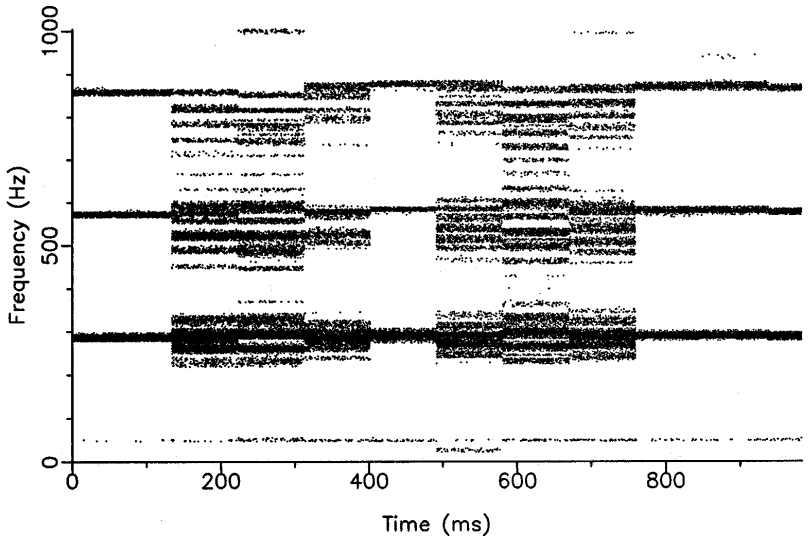


Figure 9. Narrow-band spectrogram of a patient with unilateral laryngeal nerve paralysis, corresponding to the waveform in Figure 8.

In the F_o profile, shown in Figure 10, the F_o contour again shows some large fluctuations in the segments where 30 Hz modulation takes place. The F_o extractor is trying to recognize the presence of a 285 Hz fundamental, but gets confused with the modulation frequency. The second order perturbation function again exhibits large fluctuations (much greater than $\pm$10%), indicating that perturbation measures will be unreliable. Finally, the power spectrum of the F_o contour shows the modulation frequency as a strong peak between 30 and 40 Hz.

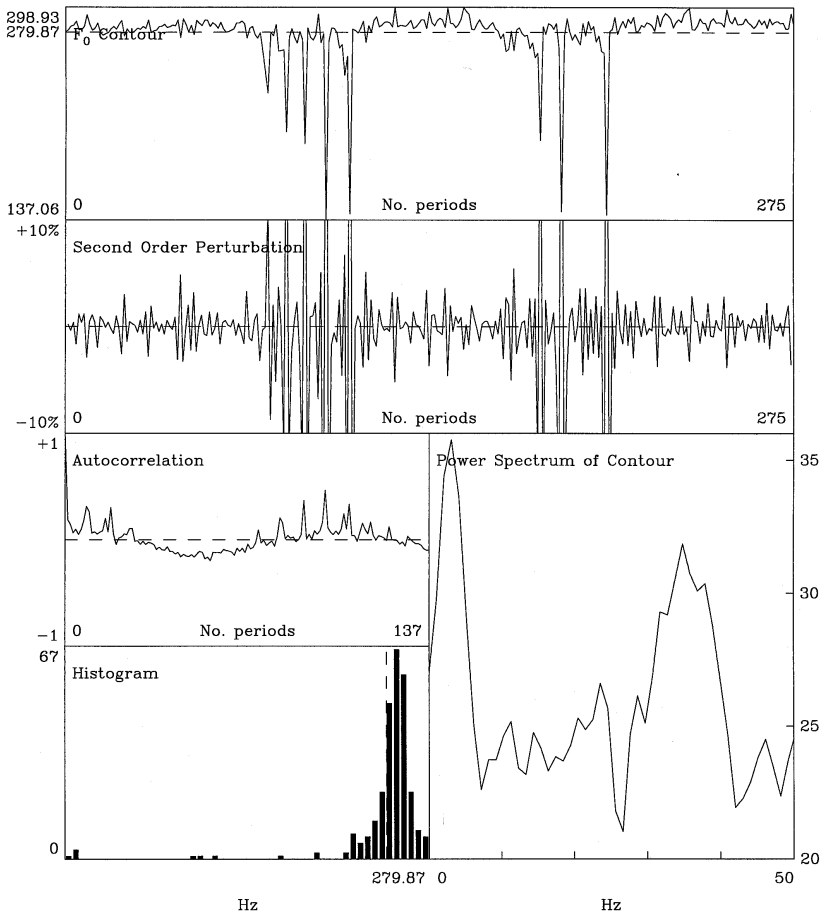
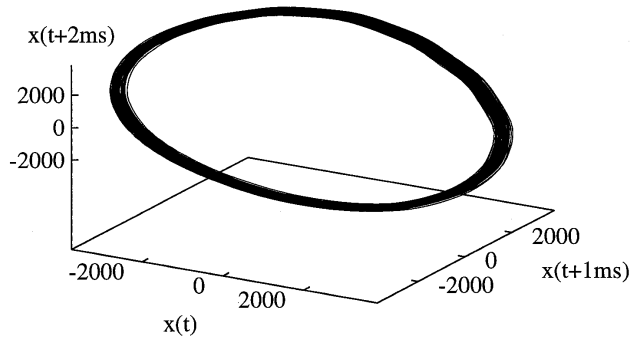
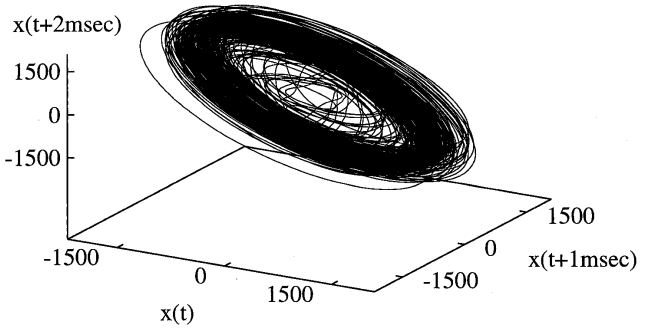


Figure 10. Fundamental frequency (F_o) profile for the patient with unliateral laryngeal nerve paralysis, corresponding to Figures 8 and 9.

A new method of analysis has recently been applied to determine the structure in complex vibrations. By examining many events in so-called *phase space* (a space that contains all of the independent variables of a system), a path can be observed to which the system is attracted. This *attractor* is the locus of points in phase space as time marches on (Figure 11). It often takes thousands of observation points before any structure can be detected. Figure 11a shows the attractor for a normal voice, whereas Figure 11b shows an attractor for the voice of the aforementioned patient with nerve paralysis. The modulations create the appearance of a torus rather than a narrow ring (a limit cycle). Interested readers in nonlinear dynamics and phase portraits are referred to Bergé et al. (1984) or Moon (1987) for an introduction to the subject. For applications of nonlinear dynamics to vocal fold vibration, the articles by Baken (1990), Herzel et al. (1991), Titze et al. (1993), Berry et al. (1994), and Herzel et al. (1994) are useful and interesting reading.



(a)



(b)

Figure 11. Phase portrait of the patient with unilateral laryngeal nerve paralysis. The microphone signal was low-pass filtered above the mean F_o and time-delayed samples were used to plot two "independent" variables.

TEST UTTERANCES

The traditional clinical goals of constructing test utterances are to determine (1) how voice effects speech intelligibility and communication effectiveness and (2) what insight can be gained about laryngeal health or general body condition. An additional pedagogical goal would be to determine (3) how the effectiveness of vocal training can be quantified.

Historically, clinicians have used a battery of test utterances that progress from vowels to isolated syllables or words to complete sentences or paragraphs. Almost everyone agrees that the tasks must reveal control of pitch, loudness, and some aspect of vocal quality. In addition, the interaction among respiratory, phonatory, and articulatory components of speech are important to most clinicians.

Table 1 shows a set of utterances. The top half of the table lists a variety of nonspeech utterances, and the bottom half lists some speech utterances. The battery includes most of the utterances used historically but expands the list significantly in the direction of dynamic testing. Phonatory *glides* are introduced for the assessment of coordinated muscle activity in the larynx and respiratory system.

All utterances may be customized to an individual's Voice Range Profile (VRP). This VRP should be obtained first to establish the bounds for further testing. Low, medium, and high pitch can then be defined as some percentage of the F_o range, say 10%, 50%, and 80%. The same can be done to define soft, medium, and loud intensity. With these definitions, sustained vowels are elicited at strategic locations within the VRP to determine phonatory stability. This is followed by [s] and [z] consonants for respiratory competence. Finally, a series of pitch, loudness, adduction, and register glides are executed to determine range, speed, accuracy, and stability in phonation. Tests of this kind were discussed by Kent et al. (1987).

In the second half of the table, speech and song material is used with increasing phonetic, emotional, and artistic complexity. After traditional counting, an all-voiced sentence is first used to test F_o control independent of adductory control. This is followed by a sentence with frequent voicing onset and offset tailored to specific larynges. The "Rainbow Passage," an often-used paragraph in speech diagnostics for English, is then administered as a *de facto* standard. At this point, some parent-child speech is attempted. Exaggerated F_o, intensity, and register patterns emerge in this test as subjects mimic typical parentese, such as those found in the "Goldilocks" story. Further testing of extreme F_o and intensity patterns (with highly expressive vocalizations) comes with a dramatic recitation, such as one of Shakespeare's soliloquies. Finally, a portion of a familiar song ("Happy Birthday") is sung in both modal and falsetto register to examine "heavy" and "light" production in a singing mode. The use of falsetto singing has been found to be useful in detecting swelling of vocal fold tissue (Bastian et al., 1990).

A major unanswered question is whether a person's ability to speak or sing can in any way be assessed with the nonspeech tasks. One would hope that wide ranges of pitch and loudness in the Voice Range Profile, for example, would predict highly expressive intonation, stress, and loudness patterns in speech, but there is no guarantee of that. For assessment of voice disorders, large inaccuracies in pitch and intensity glides should be a predictor of abnormal prosodic contours in speech, but again, this remains an open research question.

Table 1
Proposed Test Utterances

N O N S P E E C H

Voice Range Profile defines test frequencies and intensities (low = 10% of F_o range, medium = 50% of F_o range, high = 80% of F_o range; soft = 10% of intensity range, medium = 50% of intensity range, loud = 80% of intensity range)

Sustained [b], [i], [u] Vowels
1. low, soft, 2s
2. low, loud, 2s
3. high, soft, 2s
4. high, loud, 2s
5. medium high, medium loud, 2s
6. comfortable pitch and loudness, 2s
7. comfortable pitch and loudness, maximum duration

Sustained [s] Consonant
comfortable pitch and loudness, maximum duration

Sustained [z] Consonant
comfortable pitch and loudness, maximum duration

Pitch Glides
1. low-high-low, one octave, 0.25 Hz
2. low-high-low, one octave, 1.0 Hz
3. low-high-low, one octave, maximum rate

Loudness Glides
1. soft-loud-soft, 0.25 Hz
2. soft-loud-soft, 1.0 Hz
3. soft-loud-soft, maximum rate

Adductory Glides [b] and [hb]
1. onset-pressed-offset, 0.1 Hz
2. onset-pressed-offset, 2.0 Hz
3. onset-pressed-offset, maximum rate

Register Glides
1. modal-pulse-modal, 0.1 Hz
2. modal-falsetto-modal, 0.1 Hz
3. modal-falsetto-modal, maximum rate, as in yodeling

S P E E C H

Counting from 1 to 100, comfortable pitch and loudness
All voiced sentence, "Where are you going?", soft, medium, loud
Sentence with frequent voice onset and offset "The blue spot is on the key again", soft, medium, loud
Oral reading of "Rainbow Passage"
Descriptive speech, "Cookie Theft" picture
Parent-child speech, "Goldilocks and The Three Little Bears"
Dramatic speech involving deep emotions (fear, anger, sadness, happiness, disgust)
Singing part of "Happy Birthday to you", modal and falsetto register

SUMMARY OF RECOMMENDATIONS

The workshop participants discussed and approved a number of recommendations. They are divided into several subheadings dealing with classification of signals, extraction of cyclic parameters, test utterances, acquisition of signals, file formats, and data base sharing. Whenever references are given, they are not intended to be the original or most authoritative, but those that contain more detailed explanations by the workshop participants and their colleagues.

A. Classification of Signals and General Analysis Approach

A1. It is useful to classify acoustic voice signals into three types. Type 1 signals are nearly-periodic: type 2 signals contain intermittancy, strong subharmonics or modulations; type 3 signals are chaotic or random. A spectrogram, a phase portrait, or a cyclic parameter contour is useful in making the classification.

A2. For type 1 signals, *perturbation analysis* has considerable utility and reliability. As a practical guideline, perturbation measures less than about 5% have been found to be reliable (Titze & Liang, 1993).

A3. For type 2 signals, *visual displays* (e.g., spectrograms, phase portraits, or next-cycle parameter contours) are most useful for understanding the physical characteristics of the oscillating system. Perturbation measures by themselves are unreliable and contain little pattem information. Thus, assessment of voice disorders and phonatory characteristics is best accomplished on the basis of the entire visual display rather than a single measure.

A4. For type 3 signals, *perceptual ratings* of roughness (and any other auditory manifestation of aperiodicity) are likely to be the best measures for clinical assessment (Gerratt & Kreiman, 1995; Rabinov, 1995). Various system dimensions (e.g., fractal dimension, attractor dimension or Lyapunov exponent) may in time prove to be a viable acoustic compliment to perceptual ratings. Phase portraits are useful visual confirmation of high dimensionality (Herzel et al., 1994).

B. Extraction of Cyclic Parameter Contours and Perturbation Measures

B1. Since the definition of *fundamental frequency* F_o is unambiguous only for type 1 signals, any per-cycle measurement of F_o and its statistical variation (perturbation) for type 2 or type 3 signals cannot be reliably extracted.

B2. Since the definition of a *per-cycle amplitude* is based on the definition and extraction of a fundamental period ($1/F_o$), any measurement of per-cycle amplitude and its statistical variation (perturbation) for type 2 or type 3 signals cannot be reliably extracted. For type 1 signals, the per-cycle amplitude (peak value, peak-to-peak value, RMS, etc.) needs to be clearly defined because perturbation values are dependent on these definitions.

B3. A *short-term average cyclic parameter contour* (e.g., average F_o contour, average amplitude contour) is determined on the basis of a minimum cost path through a sequence of candidate cyclic parameters. The candidate cyclic parameters are derived from local "minimum distance" measures between segments of the waveform separated in time. Dynamic programming algorithms (Talkin, 1995), correlation algorithms (Milenkovic, 1987), and cepstral algorithms (Hess, 1983; 1995) are examples of this technique. For correlation F_o tracking, (1) center-clipping is not needed, (2), the cross-correlation is preferred to the autocorrelation, and (3) the confusions in F_o resulting from subharmonics is best resolved with global analysis, such as dynamic programming (Milenkovic, 1995). Average cyclic parameter contours can be extracted from both type 1 and type 2 signals, but when bifurcations in type 2 signals occur (sudden qualitative changes in the waveform), it is likely that some arbitrary decisions by the extraction algorithm will affect the contour in a non-unique way.

B4. An *event-based cyclic parameter contour* (e.g., F_o contour, amplitude contour, open quotient contour) is obtained on a per-cycle basis by marking cyclic events (peaks, zero crossings, etc.) or by making "minimum distance" measures between segments of the waveform separated by one cycle. It is often helpful to obtain a *short-term average cyclic parameter contour first* (see B3) to place candidate event markers. The event-based cyclic parameter contours are highly susceptible to error in type 2 or type 3 signals because the extraction algorithms are often dependent on specific waveform patterns. The contours can be used for visual display, but are not recommended for perturbation measures on type 3 signals. In type 1 signals, the "minimum distance" measure (also called "waveform matching"; Titze & Liang, 1993) is the most accurate extraction method and is recommended for high precision perturbation analysis.

B5. In any voice perturbation analysis, the *perturbation function* should be made clear. The *de facto* standard has been the first-order perturbation function, but when long-term trends are apparent in the cyclic parameter contour, the second-order perturbation function is recommended for elimination of these trends.

B6. In any voice perturbation analysis, the *perturbation measure* should be made clear. The *de facto* standard is the mean absolute (rectified) measure.

B7. Before applying a statistical measure to a perturbation function, it is important to study the distribution (e.g., the histogram of the cyclic parameter contour) to determine the appropriateness of the measure (Pinto & Titze, 1990; Lemke & Samawi, 1995).

B8. The use of logarithms in amplitude perturbation measures is not recommended because ratios of adjacent amplitude are small.

B9. All perturbation measures should be expressed in percent by normalizing the mean value of the cyclic parameter. Exceptions are when the mean value is zero or the parameter is time-varying (as in a glide or running speech).

B10. The length of an analysis window should be on the order of 100 cycles to obtain a stable perturbation measure (Scherer et al., in press).

C. Test Utterances for Voice Analysis

Test utterances for acoustic voice analysis can be classified as (a) sustained vowels and sustained voiced consonants, (b) vowels and voiced consonants with prescribed patterns of a cyclic parameter (e.g., glides, scales, etc.), or (c) speech utterances.

C1. Sustained vowels should continue to be used for voice perturbation analysis because they elicit a stationary process in vocal fold vibration.

C2. If utterances with prescribed patterns (e.g., F_o glides, intensity glides, etc.) are used, the patterns should be removed in the analysis and not included as part of the perturbation measure.

C3. Whenever possible, a high vowel ([i] or [V] and a low vowel ([b] or [<] should be used to report voice perturbation because source-vocal tract interactions are vowel dependent and can therefore influence laryngeal behavior.

C4. Multiple tokens of a sustained vowel (on the order of 10) are necessary to obtain reliable perturbation measures (Scherer et al, in press). Generally, the number of tokens required increases with the size of the perturbation measure.

C5. Since voice perturbations vary with F_o, intensity, and voice quality, these quantities should be defined whenever inter and intra-subject differences are reported.

D. Acquisition of Acoustic Voice Signals

D1. For type 1 signals for which a perturbation measure of the order of 0.1% is to be extracted to 10% accuracy, the following recommendations are made:

a. A professional-grade condenser microphone (omnidirectional or cardioid) with a minimum sensitivity of -60 dB should be used (Titze & Winholtz, 1993).

b. For steady vowel utterances, the mouth-to-microphone distance can be held constant and less than 10 cm (preferably 3-4 cm) in order to avoid an artificial wow and to maintain a high signal-to-noise ratio; a miniature head-mounted microphone is recommended (Winholtz & Titze, in press). This recommendation does not necessarily apply to general speech analysis, where breath noises can contaminate the signal at close distances.

c. Close microphone distances require off-axis positioning (45^o to 90^o from the mouth axis) in order to reduce aerodynamic noise from the mouth in speech.

d. The amount of room reverberation, room noise, and proximity to reflecting surfaces inside the recording booth need to be controlled. Exact recommendations are forthcoming.

e. A 16-bit A/D converter or DAT recorder is recommended, but this must be accompanied by conditioning electronics (amplifiers, filters) that have signal-to-noise ratios in the 85-95 dB range (Doherty and Shipp, 1988).

f. Sampling frequencies of 20-100 kHz should be used, depending on the degree of interpolation between samples that the analysis software provides (Titze, Horii & Scherer, 1987; Milenkovic, 1987; Deem et al., 1989).

D2. Manufacturers of workstations for acoustic voice analysis should be encouraged to provide DC coupling and low-frequency fidelity in acquisition hardware to accommodate physiologic signals (e.g., an electroglottograph, a flow mask) that augment the microphone signal. For all input signals, real-time feedback for clipping should be provided to avoid overloading the A/D converters. For DC coupling, there should be minimal drift and the drift should be reported and calibratable.

D3. Line-level inputs (on the order of a few hundred millivolts) should be provided as a direct interface to the outputs of transducers, so that expensive high fidelity analog preamplifiers can be bypassed.

D4. A digital audio tape (DAT) recorder should be used to store signals, unless A/D conversion is directly to the computer (Doherty & Shipp, 1988).

D5. Recordings should be made in a sound-treated room (ambient noise < 50 dB); given that 120 Hz is very close to the average normal male speaking F_o, special care should be given to the removal of noise sources in the room that create 60 Hz hum and its associated harmonics. In general, one should specify the spectral weighting of the allowable noise in a sound-treated room. This is particularly important if inverse filtering from the microphone signal is attempted.

E. File Formats

A number of file formats exist for speech and voice data (e.g. SPHERE, ILS, RIFF, Kay's NSP, CSRE40, CSpeech and NCVS92). These formats have been developed over many years and have a number of adherents.

E1. SPeech HEader REsources (SPHERE), developed by the National Institute of Standards and Technology (NIST), has the potential for high usage within the general scientific community, and is recommended. It is currently being used for the dissemination of the Texas Instruments-MIT-NIST (TIMIT) speech database. It contains a 1024 byte ASCII header followed by the data (which may be compressed). The header consists of a fixed format portion identifying the header type, and the length of the header. Following this is the object-oriented free format portion of the header, which describes such characteristics as sampling rate, channel count, and coding method. Software utilities have been provided by NIST for reading, writing and compressing data files. Information and software are available through Jon Fiscus, National Institute of Standards and Technology, Bldg. 225, Room A-216, Gaithersburg, Maryland 20899.

E2. If the data are to be used outside the general scientific community, or consists of multiple sources (e.g. video and audio), or requires compatibility with common PC based sound cards, the Microsoft RIFF format (which defines WAV files) is recommended. The RIFF format is very similar to Kay Elemetric's NSP format, which has been used widely in clinically-based voice laboratories. Kay provides utilities for conversion between RIFF and NSP.

E3. If neither of these formats are suitable, it is recommended that the format chosen conform to a structure in which the header and data are isolated, so that others may strip the header to gain access to the data. NCVS92, ILS, RIFF, and SPHERE are some of the formats that adhere to this principle.

F. Data Base Sharing

F1. For speech materials, there are a number of data bases available which have particular phonetic characteristics e.g., the TIMIT data base described in E1 is phonetically balanced, and uses Shibboleth sentences. Other data bases available are the Wall Street Journal (WSJ), the Resource Management (RM), and Air Transportation Information Systems (ATIS). These are just a few of the many available. They can all be obtained from the Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104, email: LDC@unagi.cis.upenn.edu, world wide web: ftp.\\www.cis.upenn.edu.

F2. Kay Elemetrics is offering a CD-ROM entitled *Disordered Voice Database of the Massachusetts Eye and Ear Infirmary Voice and Speech Lab*. This database has entries from over 700 subjects and includes both video and audio records. For more information, contact Kay Elemetrics, 2 Bridgewater Lane, Lincoln Park, NJ 07035-1488.

F3. For steady vowels and voiced consonants, vowels and consonants with dynamic characteristics such as glides, and sentences eliciting highly expressive voice production, the NCVS is currently producing its own data base. Information about this data base may be obtained from Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts, 1245 Champa Street, Denver, CO 80204.

G. Data Base Management

Data base management (attribution, classification, annotation, etc.) was not discussed in the workshop, but should be addressed in the future as a growing concern. As more databases are being created and mixed in large storage and retrieval systems, automated database indexing will become a necessity.

GLOSSARY OF TERMS

Abduction: Movement of the vocal folds in the process of separation.

Abduction Quotient: The ratio of the glottal half-width at the vocal processes to the amplitude of vibration of the vocal fold.

Adduction: Movement of the vocal folds in the process of approximation.

Amplitude: In a sinusoid, the magnitude of the maximum positive or negative excursion from the zero axis; in a complex periodic signal, the positive or negative peak, peak-to-peak, or root-mean-squared (RMS) value in a given cycle; in a voice signal, instantaneous amplitude is measured between two cyclic (recurring) events, whereas average amplitude is estimated over a series of cycles on a least error criterion.

Amplitude-to-length Ratio: The ratio of the mid-membranous amplitude of vibration to the length of the membranous vocal fold.

Aperiodicity: The absence of periodicity, or superposition of periodic oscillations with frequencies of non-integer ratios. Generally, any deviation from periodicity.

Aphonia: Absence of phonation; the inability to set the vocal folds into vibration, either constantly or intermittently; whisper is often the replacement for intended phonation.

Aspiration: The sound made by turbulent airflow preceding or following vocal fold vibration, as in [ha] or [ah].

Asthenic (Lax) Voice: A voice that appears too low in effort, weak; hypofunction of laryngeal muscles is apparent.

Attractor: A trajectory (or more strictly, an invariant set) in phase space to which a system asymptotes when stationarity is achieved.

Bifurcation: A qualitative change in the behavior of a nonlinear dynamical system when a parameter of the system is varied.

Biphonia: Phonation with two independent pitches; acoustically, there are two non-commensurate fundamental frequencies, which can appear as nonparallel harmonic lines in a spectrogram as either or both pitches change. [Theoretically, the lines may be parallel but not rationally dependent]. This definition can be extended to *triphonia* or *multiphonia*.

Bleat: See *flutter*.

Breathy Voice: Containing the sound of breathing (expiration) during phonation; acoustically, breathy voice, like falsetto, has most of its energy in the fundamental, but a significant component of noise is present due to turbulence in the glottis. In hyperfunctional breathiness, air leakage may occur in various places along the glottis, whereas in normal voice, air leakage is usually at the vocal processes.

Chaos: A qualitative description of the behavior of a dynamical system that is deterministic (nonrandom) but aperiodic.

Chest Register: A register that appears to be related to a strong phase delay between the upper and lower margins of the vocal folds; in singing, a tracheal resonance seems to enhance this register; chest register is often used interchangeably with modal register.

Convergent Glottis: The glottis narrows from bottom to top.

Covered Voice: A darkened quality obtained by rounding and protruding the lips or by lowering the larynx; the term is likely to stem from covering (fully or partially) the mouth of a brass instrument to obtain a muffled sound; acoustically, all formants are usually lowered and a stronger fundamental is obtained.

Creaky Voice: A voice that sounds like a creaking door, like two hard surfaces rubbing against each other; acoustically, a complex pattern of subharmonics and modulations is observed that reflect a complexity of modes of vibration of the vocal folds.

Crossover Frequency: The fundamental frequency for which there is an equal probability for perception of two adjacent registers.

Cyclic Parameter: Any quantity that is defined within a cycle (e.g. amplitude, period, open quotient, skewing quotient in the context of any periodic repetition of the event).

Dichrotic: See *biphonation.*

Diplophonia: Phonation in which the pitch is supplemented with another pitch that corresponds to a frequency an octave higher; some roughness is usually perceived; dynamically, there is a period doubling (an $F_o/2$ subharmonic).

Divergent Glottis: The glottis widens from bottom to top.

Dysphonic: Abnormal in phonation.

Falsetto Register: A register in which the voice is perceived to be continuous (non-pulsed) and weak in timbre; acoustically, the fundamental carries the greatest amount of energy; physiologically, only partial contact is made between the vocal folds, especially vertically.

Fluctuation: A back and forth irregular movement, usually indicating instability in a system.

Flutter: Phonation with amplitude or frequency modulations (or both) in the 8-12 Hz range; physiologically; also called bleat, as the bleating of a lamb.

Forced Oscillation: Oscillation imposed on a system by an external periodic source.

Free Oscillation: An oscillation without any imposed driving forces.

Frequency: The number of events per second; in a sinusoid, the number of cycles (2π radians) per second.

Fundamental Period: In a periodic signal, the smallest value T_o that satisfies the relation $f(t+T_o)=f(t)$ for all time t; in a voice signal, instantaneous T_o is the time between two cyclic (recurring) events, whereas average T_o is the smallest constant inter-event duration that best matches a series of prominent recurring events.

Fundamental Pitch: In a voiced sound, the lowest perceived pitch associated with vocal fold vibration.

Fundamental Frequency: The inverse of fundamental period.

Glottalized Voice: A voice that contains frequent transient sounds (clicks) that result from relatively forceful adduction or abduction during phonation.

Glottis: The airspace between the vocal folds.

Harmonic Frequencies: Frequencies that are related to the fundamental frequency by an integer ratio.

Histogram: A display of the number of times a variable takes on a certain value, or a small range of values, in its total range; also known as the distribution density of the variable.

Hoarse Voice: The combination of rough voice and breathy voice.

Honky (Nasal) Voice: A voice quality associated with the excessive acoustic energy coupling to the nasal tract; acoustically, nasality is characterized by a low-frequency murmur and spectral zeros.

Jitter: A short-term (cycle-to-cycle) variation in the fundamental frequency of a signal.

Lift: A transition point along a pitch scale where vocal production becomes easier (lifted). The term is used to describe register transitions.

Loft: A suggested term for the highest (loftiest) register; usually referred to as falsetto voice.

Loudness: The psychoacoustic perceptual measure of a sound on a strong-weak continuum; the primary acoustic correlate is sound pressure level.

Mean: The value obtained by adding up N numbers and dividing by N.

Mean Rectified: The value obtained by first rectifying (taking the absolute value of) a set of numbers and then taking the mean.

Median: The value obtained by working a histogram of a set of numbers and letting the number of entries above and below the value be equal.

Median Rectified: The value obtained by first rectifying (taking the absolute value of) a set of numbers and then finding the median.

Modal Register: A register that appears to be related to a strong phase delay between the upper and lower margins of the vocal folds; auditorily, contact is made between the vocal folds during the closed phase, both vertically and horizontally; the voice is perceived to be continuous (non-pulsed) and relatively rich in timbre; acoustically, the spectral slope of the glottal source (volume velocity) waveform is on the order of 12-15 dB/octave.

Mode (of Vibration): A characteristic spatial pattern of vibration that can (in principle) exist in isolation, but ordinarily forms a building block (together with other modes) for complicated vibrating patterns.

Modulation: The systematic variation of a cyclic parameter (e.g. amplitude or fundamental frequency) over several cycles of phonation.

Nasal Voice: Associated with excessive opening of the velar port in vowel production; see honky voice and twangy voice.

Natural Oscillation: Oscillation without imposed driving forces; usually observed after an impulse of energy is given to a system.

Oscillation: A repeated back and forth movement, particularly when self-sustained (see self-sustained oscillation).

Passaggio: Passages on a pitch scale where the voice tends to change register involuntarily.

Period Doubling: A bifurcation in which two adjacent cycles become unequal, but together form a new period of twice the original length.

Periodicity: The property of a time series such that f(t+nT)=f(t), where T is the period and n is any positive integer.

Perturbation: A disturbance, or small change, in a cyclic variable (period, amplitude, open quotient, etc.) that is constant in regular periodic oscillation.

Perturbation Function: A time series of differences between selected cyclic parameters that are delayed or advanced in time (e.g., the first-order difference function of the F_o contour).

Perturbation Measure: An average value of the perturbation function over an analysis window of several cycles.

Phase Space: A space defined by two or more independent dynamical variables (in particular, position and velocity) to plot the trajectory of a dynamically varying object.

Phonation: The process of creating sound by vocal fold vibration.

Pitch: The psychoacoustic perceptual measure of a sound on a high-low continuum; the primary acoustic correlate is fundamental frequency.

Pressed Voice: Phonation in which the vocal processes of the arytenoid cartilages are pressed together, resulting in a constricted glottis with relatively low airflow; there is also medial compression of the vocal fold tissue; acoustically, the fundamental is weakened relative to the overtones.

Pulsed Phonation: Phonation in which temporal gaps are perceived; acoustically, energy "packets" are perceived below about 70 Hz, where formant energy effectively dies out prior to re-excitation with a new glottal pulse; pulsed phonation or pulse register is also called vocal fry, apparently because of its similarity with popping sounds that are emitted from a hot frying pan.

Rectification: The process of taking the absolute value of a function or time series (i.e., making all negative values positive).

Register: A major category of voice quality (e.g., modal, falsetto, pulse, chest, head, whistle).

Resonant Voice: A voice quality that rings on, "carries" well; acoustically, ample formant energy is excited.

Ringing (Resonant) Voice: A brightened quality, apparently obtained by enhanced epilaryngeal resonance, which produces a strong spectral peak around 2500-3500 Hz. In effect, there is a clustering of the formants F_3, F_4 and F_5; the combined resonances are often called the "singer's formant".

Root-mean-squared: The operation that involves first squaring each of a set of numbers, then finding the mean value of the squared numbers, and finally taking the square root of the mean value.

Rough Voice: An uneven, bumpy quality that appears to be unsteady in the short-term, but stationary in the long-term; acoustically, the waveform is often aperiodic, with the modes of vibration lacking synchrony, but voices with subharmonics can also be perceived as rough.

Self-Sustained Oscillation: An oscillation that continues indefinitely without a periodic driving force; since the net energy loss per cycle must be zero, self-sustained oscillation requires an energy source.

Shimmer: A short-term (cycle-to-cycle) variation in the amplitude of a signal.

Spectral Slope: A measure of how rapidly energy decreases with increasing frequency, or, for periodic wave forms, with increasing harmonic number. Also known as *spectral tilt* or *spectral roll-off*.

Stationarity: The property of a signal that suggests no long-term drifts; the autocorrelation function $<x(t) * x(t+\delta)>$ depends only on δ, not on t, and decays to zero with increasing t; the spectrogram remains constant over time.

Strained (Tense) Voice: A voice that appears effortful; visually, hyperfunction of the neck muscles is apparent; the entire larynx seems compressed.

Strohbass: Literal translation from German, "straw bass", because of its perceptual similarity to crackling straw; it is effectively the pulse register when used in singing.

Subharmonic Frequencies: Frequencies that lie between or below the harmonic frequencies and are rational divisions of the fundamental frequency (e.g. 1/2, 1/3) or their integer multiples.

Temporal Gap Transition: The transition from a continuous sound to a series of pulses in the perception of vocal registers.

Tremor: A 1-15 Hz modulation of a cyclic parameter (e.g. amplitude or fundamental frequency), either of a neurologic origin or an interaction between neurological and biomechanical properties of the vocal folds. See *flutter*, *vibrato*, and *wow*.

Trill: A rapid alternation of a primary note with a secondary note (usually a semitone or a tone higher); used as an ornament in music.

Trillo: A rapid repetition of the same note in the 8-12 Hz range; used as an ornament in music.

Twangy Voice: A sharp, bright quality, as produced by a plucked string. Twang is often attributed to nasality, but it is probably more laryngeally-based. It is often part of a dialect or singing style.

Variability: Literally, the ability of something to vary, by design or by accident. More formally, the amount of variation as determined by a statistical measure.

Ventricular Phonation: Phonation with the false vocal folds; unless intentional, it is generally considered an abnormal muscle pattern dysphonia associated with hyperactivity in the false fold region.

Vibrato: A natural ingredient of a singing voice, especially in classical Western singing; acoustically, a 4-7 Hz sinusoidal modulation of F_o and/or intensity; the modulation extent is typically $\pm$3% in frequency, but varies considerably in amplitude. Physiologically, the origin of natural vibrato lies in laryngeal muscle contraction rather than lung pressure modulations.

Whisper: Speech produced by turbulent glottal airflow in the absence of vocal fold vibration.

Whistle Register: A register in which the sound is perceived as a whistle, usually high in pitch and flute-like in quality; physiologically, the claim is that a posterior glottal gap can serve as an orifice for vortex shed ding and an epilaryngeal resonator can reinforce the sound, but the resonance mechanism is yet speculative.

Wobble: See *wow*.

Wow (Wobble): Phonation with amplitude and/or frequency modulations in the 1-3 Hz range.

Yawny Voice: A quality associated with a lowered larynx and widened pharynx, as in a yawn.

Acknowledgement

The author has been greatly influenced by the writings of (and personal communication with) Dr. Hanspeter Herzel. He read the manuscript with interest and care and made many suggestions.

REFERENCES

Aronson, A., Ramig, L., Winholtz, W., & Silber, S. (1992). Rapid voice tremor, or "flutter", in amyotrophic lateral sclerosis. Annals of Otology, Rhinology & Laryngology, 101(6), 511-518.

Atal, B., Miller, J., & Kent, R. (1991). Papers in Speech Communication: Speech Processing. Woodbury, NY: Acoustical Society of America.

Baken, R. J. (1990). Irregularity of vocal period and amplitude: A first approach to the fractal analysis of voice. Journal of Voice, 4(3), 185-197.

Bastian, R. W., Keidar, A., & Verdolini-Marston, K. (1990). Simple vocal tasks for detecting vocal fold swelling. Journal of Voice, 4(2), 172-183.

Bendat, J., & Piersol, A. (Eds.). (1986). Random Data: Analysis and Measurement Procedures. New York: John Wiley and Sons.

Bergé, P., Pomeau, Y. & Vidal, C. (1984). Order Within Chaos: Toward A Deterministic Approach to Turbulence. New York: John Wiley & Sons.

Berry, D., Herzel, H., Titze, I.R., & Krischer, K. (1994). Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. Journal of the Acoustical Society of America, 95(6), 3595-3604.

Cox, N. (1989). Technical considerations in computations of spectral harmonics-to-noise ratio for sustained vowels. Journal of Speech and Hearing Research, 32(1), 203-218.

Deem, J.F., Manning, W.H., Knack, J.V., & Matesich, J.S. (1989). The automatic extraction of pitch perturbation using microcomputers: Some methodological considerations. Journal of Speech and Hearing Research, 32, 689-697.

Doherty, E., & Shipp, T. (1988). Tape recorder effects on jitter and shimmer extraction. Journal of Speech and Hearing Research, 31, 485-490.

Gerratt, B.R., & Kreiman, J. (1995). The utility of acoustic measures of voice quality. In D. Wong (Ed.), Workshop on Acoustic Voice Analysis. Iowa City, IA: National Center for Voice and Speech.

Hakes, J., Doherty, E., & Shipp, T. (1990). Trillo rates exhibited by professional early music singers. Journal of Voice, 4(4), 305-308.

Hays, W. (1988). Statistics, 4th Ed. New York: Holt, Rinehart & Winston, Inc.

Herzel, H., Steinecke, I., Mende, W., & Wermke, K. (1991). Chaos and bifurcations during voiced speech. In E. Mosekilde (Ed.), Complexity, Chaos, and Biological Evolution (pp 41-50). New York: Plenum Press.

Herzel, H., Berry, D., Titze, I.R., & Saleh, M. (1994). Analysis of vocal disorders with methods from nonlinear dynamics. Journal of Speech and Hearing Research, 37(5), 1001-1007.

Hess, W. (1983). Pitch Determination of Speech Signals: Algorithms and Devices. Berlin, Heidelberg, New York, Toronto: Springer-Verlag.

Hess, W.J. (1995). Pitch determination of speech signals - with special emphasis on time domain methods. In D. Wong (Ed.), Workshop on Acoustic Voice Analysis. Iowa City, IA: National Center for Voice and Speech.

Hillenbrand (1987). A methodological study of perturbation and additive noise in synthetically generated voice signals. Journal of Speech and Hearing Research, 30, 448-461.

Kasuya, H., Ogawa, S., Mashima, K., & Ebihara, S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. Journal of the Acoustical Society of America, 80, 1329-1334.

Kent, R. D., Kent, J. F., & Rosenbek, J. C. (1987). Maximum performance tests of speech production. Journal of Speech and Hearing Disorders, 52, 367-387.

Kent, R., Atal, B., & Miller, J. (1991). Papers in Speech Communication: Speech Production. Woodbury, NY: Acoustical Society of America.

Klingholz, F. (1987). The measurement of the signal-to-noise ratio (SNR) in continuous speech. Speech Communication, 6, 15-26.

Koda, J. & Ludlow, C. (1992). Evaluation of laryngeal muscle activation in patients with voice tremor. Otolaryngology - Head and Neck Surgery, 107(5), 684-696.

Koike, Y. (1973). Application of some acoustic measures for the evaluation of laryngeal dysfunction. Stud. Phonol., VII, 17-23.

Lemke, J., & Samawi, H.M. (1995). Establishment of normal limits for speech characteristics. In D. Wong (Ed.), Workshop on Acoustic Voice Analysis. Iowa City, IA: National Center for Voice and Speech.

Lieberman, P. (1961). Perturbations in vocal pitch. Journal of the Acoustical Society of America, 33, 597-602.

Lieberman, P. (1963). Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. Journal of the Acoustical Society of America, 35, 344-353.

Markel, J.D. & Gray, A.H., Jr. (1976). Linear Prediction of Speech. New York: Springer-Verlag.
Milenkovic, P. (1987). Least mean square measures of voice perturbation. Journal of Speech and Hearing Research, 30, 529-538.
Milenkovic, P.H. (1995). Rotation-based measure of voice aperiodicity. In D. Wong (Ed.), Workshop on Acoustic Voice Analysis. Iowa City, IA: National Center for Voice and Speech.
Miller, J., Kent, R., & Atal, B. (1991). Papers in Speech Communication: Speech Perception.Woodbury, NY: Acoustical Society of America.
Moon, F.C. (1987). Chaotic Vibrations: An Introduction for Applied Scientists and Engineers. New York: John Wiley & Sons.
Niimi, S., Horigughi, S., Kobayashi, N., & Yamada, M. (1988). Electromyographic study of vibrato and tremolo in singing. In O. Fujimura (Ed.), Voice Production, Mechanisms and Functions (pp. 403-414). New York: Raven Press.
Orlikoff, R. (1990). Heartbeat-related fundamental frequency and amplitude variation in healthy young and elderly male voices. Journal of Voice, 4(4), 322-328.
Perkell, J.S. & Klatt, D.H. (1986). Invariance and Variability in Speech Process. Hillsdale, NJ: Lawrence Earlbaum Associates, Publishers.
Pinto, N. & Titze, I. (1990). Unification of perturbation measures in speech analysis. Journal of the Acoustical Society of America, 87(3), 1278-1289.
Qi, Y.Y., Weinberg, B., Bi, N., & Hess, W.K. (1995). Minimizing the effect of period determination on the computation of amplitude perturbation in voice. In D. Wong (Ed.), Workshop on Acoustic Voice Analysis. Iowa City, IA: National Center for Voice and Speech.
Qi, Y. (1992). Time normalization in voice analysis. Journal of the Acoustical Society of America, 92, 2569-2576.
Rabiner, L.R., & Schafer, R.W. (1978). Digital Processing of Speech Signals. Englewood Cliffs NJ: Prentice-Hall.
Rabinov, C.R., Kreiman, J., & Gerratt, B.R. (1995). Comparing reliability of a perceptual and acoustic measures of voice. In D. Wong (Ed.), Workshop on Acoustic Voice Analysis. Iowa City, IA: National Center for Voice and Speech.
Ramig, L. & Shipp, T. (1987). Comparative measures of vocal tremor and vocal vibrato. Journal of Voice, 1(2), 162-167.
Rothenberg, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. Journal of the Acoustical Society of America, 53(6), 1632-1645.
Scherer, R., Vail, V., & Guo, C. (in press). Required number of tokens to establish reliable voice perturbation values. Journal of Speech and Hearing Research.
Takahashi, H., & Koike, Y. (1975). Some perceptual dimensions and acoustic correlates of pathological voices. Acta Otolaryngologica (Stockholm), Suppl. 338, 2-24.
Talkin, D. (1995). Cross correlation and dynamic programming for estimation of fundamental frequency. In D. Wong (Ed.), Workshop on Acoustic Voice Analysis. Iowa City, IA: National Center for Voice and Speech.
Terhardt, E. (1974). On the perception of periodic sound fluctuations (roughness). Acustica, 30, 201-213.
Titze, I.R., Horii, Y., & Scherer, R.C. (1987). Some technical considerations in voice perturbation measurements. Journal of Speech and Hearing Research, 30, 252-260.
Titze, I.R. (1991). A model for neurologic sources of aperiodicity in vocal fold vibration. Journal of Speech and Hearing Research, 34, 460-472.
Titze, I., Baken, R. & Herzel, H. (1993). Evidence of chaos in vocal fold vibration. In I. Titze (Ed.), Vocal Fold Physiology: Frontiers in Basic Science (pp 143-188). San Diego: Singular Publishing Group.
Titze, I. & Liang, H. (1993). Comparison of F_o extraction methods for high precision voice perturbation measurements. Journal of Speech and Hearing Research, 36(6), 1120-1133.
Titze, I.R., & Winholtz, W.S. (1993). The effect of microphone type and placement on voice perturbation measurements. Journal of Speech and Hearing Research, 36(6), 1177-1190.
Titze, I.R., Solomon, N.P., Luschei, E.S., & Hirano, M. (1994). Interference between normal vibrato and artificial stimulation of laryngeal muscles at near vibrato rates. Journal of Voice, 8(3), 215-223.
Yumoto, E., Gould, W. J., & Baer, T. (1982). The harmonics-to-noise ratio as an index of the degree of hoarseness. Journal of the Acoustical Society of America, 71, 1544-1550.
Wendahl, R.W. (1966). Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness. Folia Phoniatrica, 18, 99-108.
Winholtz, W., & Titze, I. (in press). Miniature head mount microphone for acoustic analysis. Journal of Speech and Hearing Research.