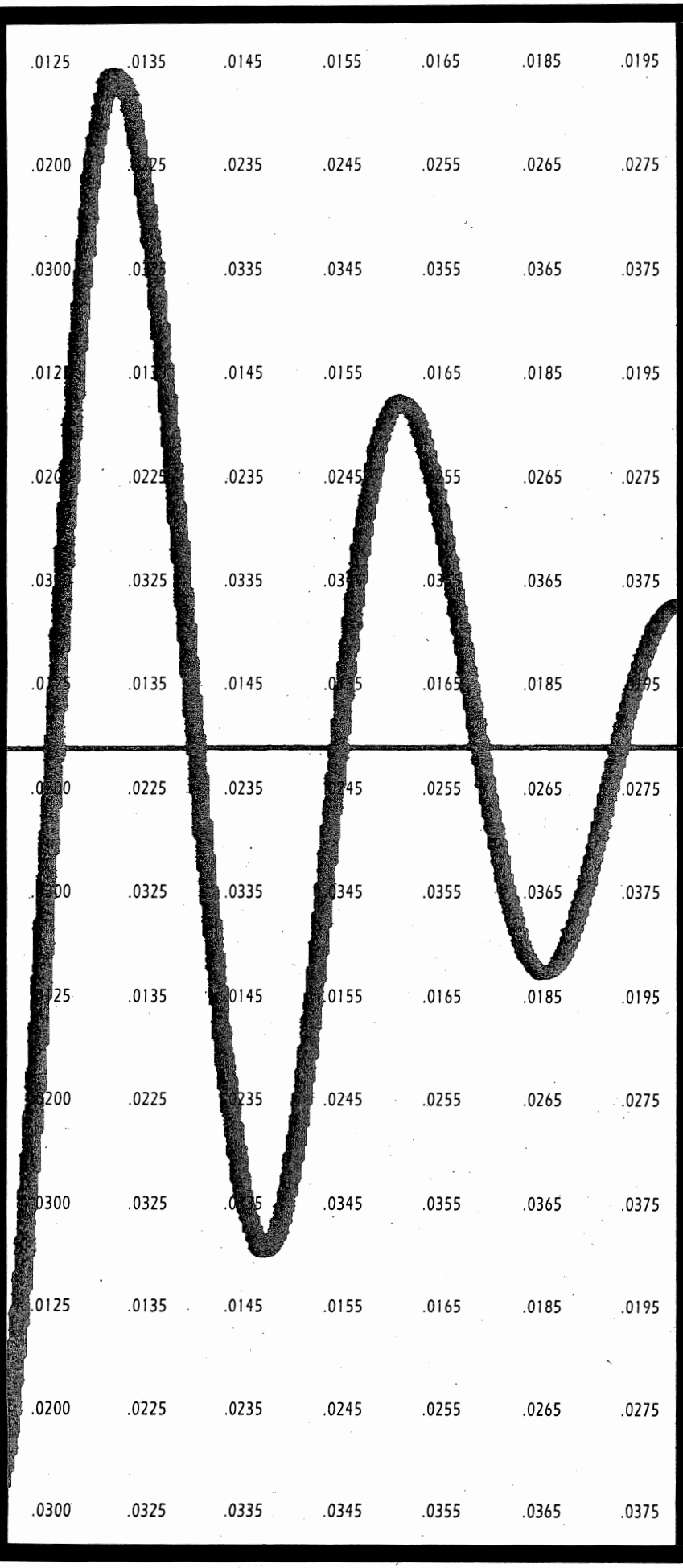


Workshop on Acoustic Voice Analysis

PROCEEDINGS

EDITED BY D. WONG

Sponsored by the

National Center for Voice and Speech
for the
National Institute on Deafness and
Other Communication Disorders

Conference Site:

Wilbur James Gould Voice Research Center
A Division of The Denver Center
for the Performing Arts

Workshop on Acoustic Voice Analysis

PROCEEDINGS

BY DARRELL WONG, 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.

FOREWORD

To understand speech and voice production and perception, scientists have traditionally studied the acoustic signal captured by a microphone. In the health sciences, the human voice has been studied as a way of revealing information about the health of an individual. While there has been considerable progress in the analysis of speech and voice signals for the diagnosis and documentation of vocal disorders, some concern has been expressed regarding the need to reach a consensus on the utility, feasibility, and standardization of voice perturbation analysis methods.

The workshop proceedings are intended to be a step toward this process. A Workshop on Acoustic Voice Analysis was held on the 17th and 18th of February, 1994, in Denver, Colorado. The site was the Wilbur James Gould Voice Research Center (then known as The Recording and Research Center), a division of The Denver Center for the Performing Arts (DCPA). Sponsorship and financial support was provided by the National Center for Voice and Speech (NCVS), and the DCPA. The NCVS is a research and training center funded by the National Institute on Deafness and Other Communication Disorders.

The Proceedings consist of written versions of topics discussed during the workshop. Some of the papers may be found in other journals. As a rule, previously published material was accepted, since the workshop was seen as a summary as much as a venue for new ideas.

Attendance and contributions were by invitation, so that a broad spectrum of the voice analysis community could be represented. While the audience and list of papers does not exhaustively represent the community, we were able to present perspectives from industry representatives, speech clinicians, speech science academicians, and medical personnel.

The topics presented include recording techniques, file formats, perturbation statistics extraction algorithms, nomenclature and classification, and the nature of perturbation. As a result, the papers ranged in style from technical summaries and algorithm descriptions to perspectives and commentaries.

As part of this Foreword, mention should be made regarding the organization of the manuscript. Each paper is identified using the first few letters of the primary author's name. The first paper, HESS, by Dr. Wolfgang Hess is the keynote address for the Workshop. He was given the task of introducing the concepts involved in pitch determination - generally considered the heart of perturbation analysis. Because perturbations are viewed as deviations from the steady state, the demarcation of fundamental periods is crucial. The next four papers (identified as TALK, MIL1, DEL, and QI), discuss the topics of pitch (or F0) marking, pitch perturbation, and amplitude perturbation. Following that are three papers (RAB, GER, and

LEM), which discuss the utility of perturbation measures and the statistical methodology for defining the 'normal limits' of speech characteristics.

The paper EPN discusses a method of pitch marking, jitter measurement, and their results as applied to aged voices. In JIA and HUA, protocols and observations are made regarding the capturing of voice samples in the context of reducing subject frequency and intensity variability. In KHE, a discussion and demonstration of new methods in spectral estimation are presented, while WON presents a qualitative discussion on the sources of perturbation from a biomechanical perspective. The latter paper was not presented during the workshop, but it has been submitted by the editor as a relevant topic. The short summary in MILD makes suggestions on hardware selection in the context of different types of voice processing. MIL2 and CUR discuss file formats, while WINH discusses microphone selection and placement as it affects perturbation measurements.

Finally, Dr. Ingo Titze has written a summary statement (TITZE), the first part of which may be considered as his personal perspective on the analysis, nomenclature and classification of voice data. The second part of the statement is a set of recommendations and a glossary of terminology. Only the recommendations should be viewed as majority opinion. The summary statement can be obtained as a separate document from the NCVS.

The proceedings have focused on a very narrow set of issues which are important to the voice analysis community. We hope that the results are informative, and that our efforts will at least generate discussion, if not a consensus, in the community.

Darrell Wong,
June, 1995.

Proceedings of the Workshop on
Acoustic Voice Analysis

February 17th and 18th, 1994
Denver, Colorado

TABLE OF CONTENTS

For more information or additional copies of this report, contact:

National Center for Voice and Speech
Wendell Johnson Speech & Hearing Center
The University of Iowa • Iowa City, Iowa 52242 • 319/335-6600

NCVS Workshop on Voice Analysis, The Center of Performing Arts, Denver,
February 1994

Pitch Determination of Speech Signals – with Special Emphasis on Time-Domain Methods

Wolfgang J. Hess

Institute for Communications Research and Phonetics (IKP), University of Bonn
Poppelsdorfer Allee 47, D–53115 Bonn, Germany
wgh@uni–bonn.de

Abstract. This paper presents a survey of methods for pitch determination of speech signals with special emphasis on time-domain methods. As speech is a time-variant signal the result of the measurement will depend on the method applied. This implies that we first define what is subsumed under the term *pitch*. From the point of view of speech production this is *rate of vocal fold vibration* or the *duration of individual laryngeal excursion cycles* which is measured in the time domain by algorithms that are able to track the signal period by period. From a more signal-oriented point of view where the emphasis is laid on periodicity of voiced speech signals, this will be *fundamental period (duration)*, or, if the measurement is carried out in the frequency domain, *fundamental frequency.* Pitch determination algorithms (PDAs) which follow this definition usually operate on the basis of some short-time, i.e., frame-to-frame representation. After a short review of these PDAs a survey of time-domain algorithms is presented. These include methods such as structural analysis of the speech signal with or without preprocessing, determination of individual periods from the first partial of the signal, determination of the point of glottal closure, and multi-channel approaches. Some remarks on glottal inverse filtering are added. The paper then discusses the issue of error analysis. Errors in pitch determination are classified into gross errors and measurement inaccuracies, and it is a main problem for any algorithm, when it detects an estimate that seems to be wrong, to detect reliably whether this is due to a measurement failure or to a momentary irregularity of the signal. The paper also addresses the possibility to use an instrument that directly measures the laryngeal excitation, notably a laryngograph, for gaining reference contours from which the PDAs can be evaluated or trained.

Pitch, i.e., *fundamental frequency* (or *rate of vocal-fold vibration*) F_0 as well as *fundamental period* T_0 takes on a key position in the acoustic speech signal. The prosodic information of an utterance is predominantly determined by this parameter. The ear is by an order of magnitude more sensitive to changes of fundamental frequency than to changes of other speech signal parameters (Flanagan and Saslow, 1958, Klatt, 1973; Harris and Umeda, 1987). The quality of vocoder speech as well as of synthetic speech (when natural-speech units are used) is essentially influenced by the quality and faultlessness of the pitch measurement (Gold, 1977). Hence the importance of this parameter claims for good and reliable measurement methods.

Besides voicing determination, pitch determination is one of the two subproblems of *voice source analysis.* In voiced speech, the vocal cords vibrate in a quasi-periodic way.

Speech segments with voiceless excitation are generated by turbulent air flow at a constriction or by the release of a closure in the vocal tract. The parameters we have to determine in voice source analysis are the *manner of excitation,* i.e., the presence of a *voiced excitation* and the presence of a *voiceless excitation,* a problem which is referred to as *voicing determination,* and – for the segments of the speech signal where a voiced excitation is present – *pitch determination.*

Automatic pitch determination has a rather long history which goes back even beyond the times of vocoding (e.g. Grützmacher and Lottermoser, 1937). The most important developments leading to today's state of the art were made in the sixties and seventies; these methods that are reviewed in this paper are extensively discussed in (Hess, 1983). Since then, few absolutely new principles have been invented; a number of methods, however, were improved and refined, whereas other solutions were revived that required an amount of computational effort appearing unrealistic at the time the algorithm was first developed. On the other hand, new techniques such als neural networks or – even more recently – the wavelet transform initiated new developments especially in time-domain pitch determination where further improvements are to be expected in the near future. With speech

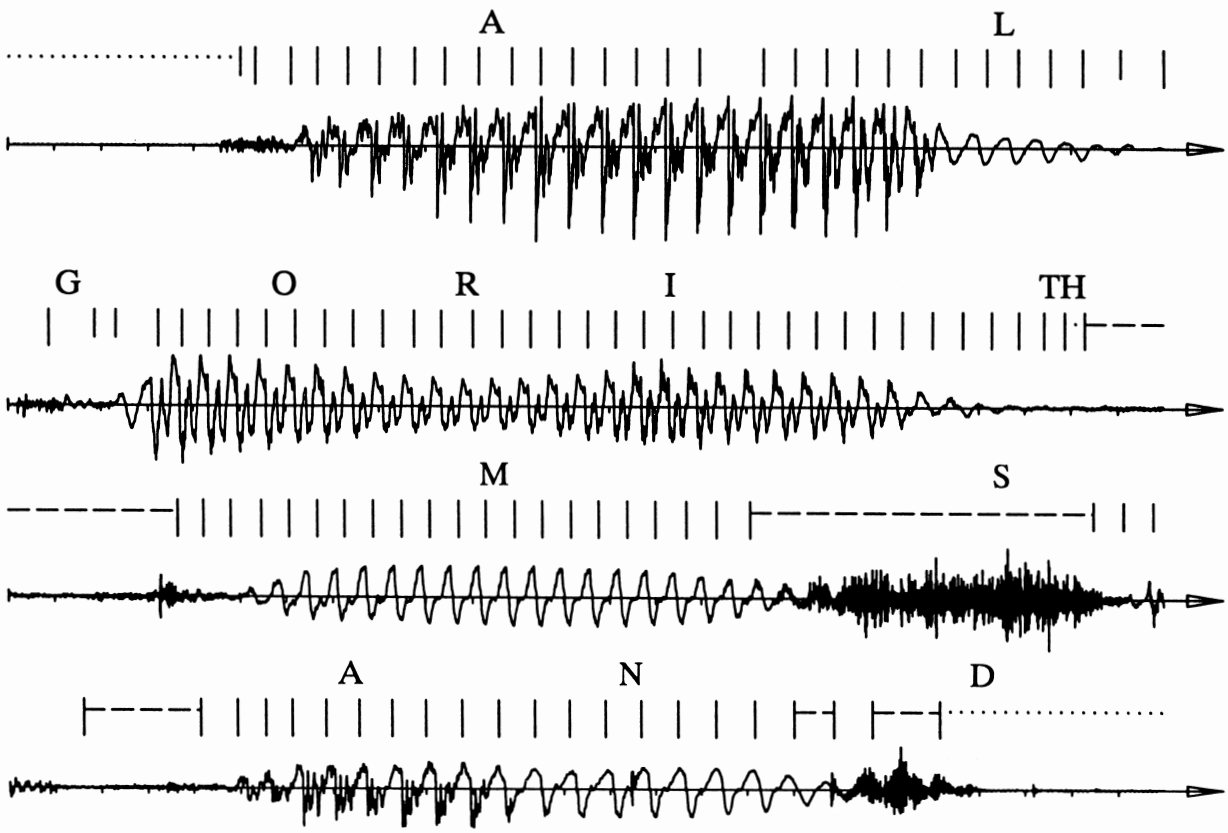


Fig. 1. Example of a speech signal (after pitch determination). Beginning of the utterance "Algorithms and devices for pitch determination". Speaker: male; undistorted signal. Scale: 250 ms per line. The analysis was done using the algorithm by Hess (1979). (– – – – – –) Voiceless, (·····) Pause, (|||||||) pitch period boundaries ("markers"). Markers indicated by short lines were found irregular

corpora coming into use that contain many labeled and processed speech data, researchers nowadays tend toward thoroughly examining and checking the performance of their algorithms.

At the first glance the task looks simple: one has just to detect the fundamental frequency of a quasi-periodic signal. Dealing with speech signals, however, the assumption of quasi-periodicity is often far from reality. Figure 1 shows an arbitrary (but typical) example of a speech signal. For a number of reasons, the task of pitch determination must be counted among the most difficult problems in speech analysis.

1) In principle, speech is a nonstationary process; the momentary position of the vocal tract may change abruptly at any time. This leads to drastic variations in the temporal structure of the signal, even between subsequent pitch periods.

2) Due to the flexibility of articulatory gestures and the wide variety of voices, there exist a multitude of possible temporal structures. Narrow–band formants at low harmonics (especially at the second or third harmonic) are a particular source of trouble.

3) For an arbitrary speech signal uttered by an unknown speaker, the fundamental frequency can vary over a range of almost four octaves (50 to 800 Hz). Especially for female voices, F_0 thus often coincides with the first formant (the latter ranging from about 200 Hz to 1400 Hz). This causes problems when inverse filtering techniques are applied.

4) The excitation signal itself is not always regular (see Fig. 2). Even under normal conditions, i.e., when the voice is neither hoarse nor pathologic, the glottal waveform exhibits occasional irregularities (Dolansky and Tjernlund, 1968; Fujimura, 1968; Lieber-

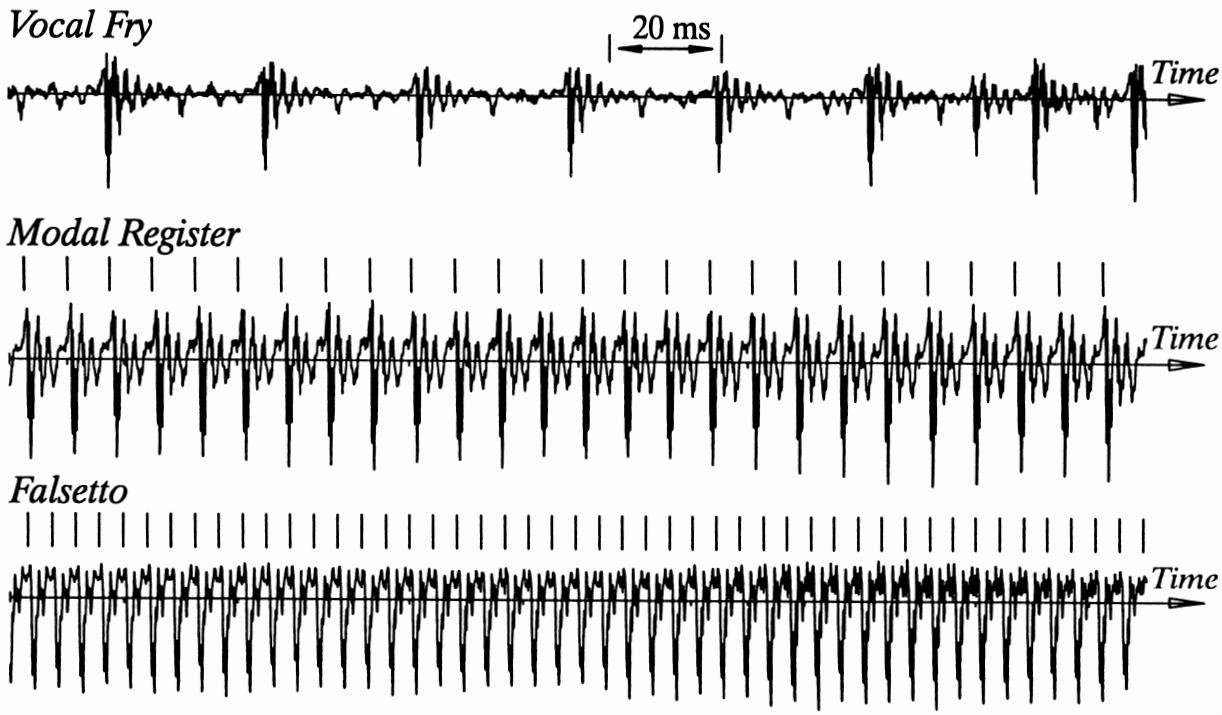


Fig. 2. Speech signals in different phonation: vocal fry (upper line); modal register, i.e., normal speech (lower line, with pitch period delimiters). Signal: sustained vowel [ɛ], speaker WGH (male)

man, 1963). In addition, the voice may temporarily fall into vocal fry or laryngealization (Hollien, 1974) which is a nonpathologic mode of voice excitation with rather large and irregular intervals between subsequent glottal pulses . Such laryngealizations are deliberately used by many speakers as boundary signals or substitutions for glottal stops (Huber, 1988) and may therefore occur anywhere in fluent speech.

5) Additional problems arise in speech communication systems where the signal is often distorted or band limited (for instance, in the telephone channel). This may be detrimental for some applications. For voice quality measurement or vocal jitter determination, for instance, even the inevitable phase distortions introduced by an ordinary analog tape or cassette recorder (cf. Fig. 3) may be intolerable.

Literally hundreds of methods for pitch determination have been developed. This paper will give a survey of the prevailing principles and discuss selected methods in more detail. First, we will deal with possible definitions of the parameter *pitch* itself (Sect. 1), followed by a gross categorization of the various principles of its determination (Sect. 2). After that we will go into a more detailed discussion of individual principles and individual solutions. Section 3 will present a brief survey of short-term analysis methods, and then Sect. 4 will deal more extensively with time-domain methods. Sections 5 and 6 will finally discuss problems of error analysis and evaluation, accurate voicing and pitch determination using instruments such as the laryngograph, and various applications.

As to the realization, we will not distinguish between a hardware device (whether analog or digital) and an algorithmic solution: they are all regarded as *pitch determination algorithms* (PDAs). In addition we will separate the problems of pitch determination and voicing determination although the two are often realized within the same algorithm; we will assume in the following that a voiced/unvoiced decision has been done already, and that there are only voiced signals to be processed by the respective PDA.

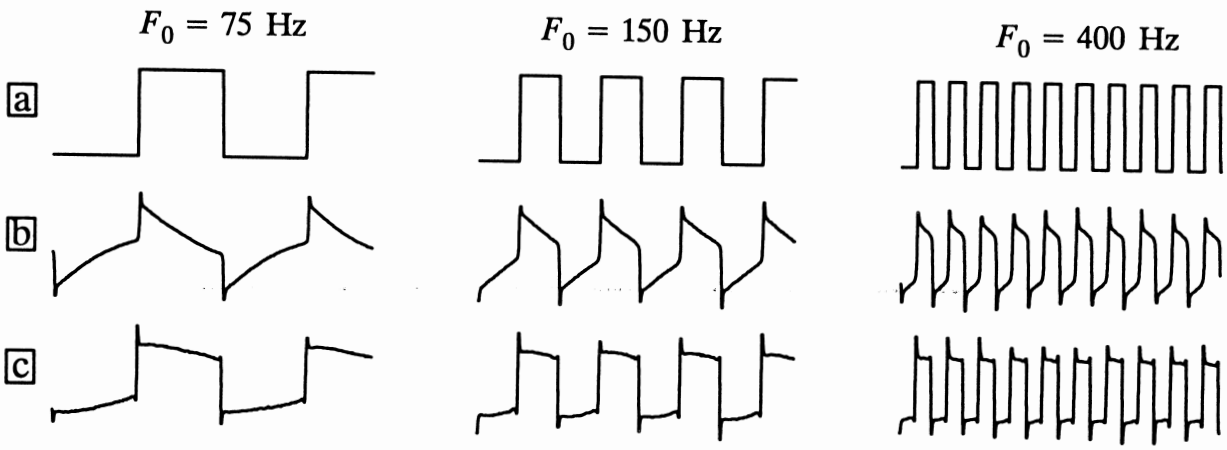


Fig. 3a-c. Phase distortions in analog tape recordings. (**a**) Rectangular waveform; (**b**) same waveform after recording on a high-quality analog tape recorder; (**c**) same waveform after rerecording it with the same recorder in backward direction

1. Basic Definitions of Pitch

Pitch can be measured in many ways. If the signal is completely stationary and periodic, all these strategies – provided they operate correctly – lead to identical results. Since the speech signal is nonstationary and time variant, however, aspects of strategy such as the starting point of the measurement, the length of the measuring interval, the way of averaging (if any), or the operating domain (time, frequency, lag etc.) of an individual algorithm start influencing the results and may lead to estimates that differ from algorithm to algorithm even if all these results are "correct" and "accurate." Before entering a discussion on individual methods, we must therefore have a look at the parameter *pitch* and provide a clear definition of what should be measured and what is actually measured.

A word on terminology first. There are three points of view for looking at a speech processing problem (Zwicker et al., 1967): the *production,* the *signal-processing,* and the *perception* point of views, respectively. In the actual case of pitch determination the production point of view is obviously oriented toward the generation of the excitation signal in the larynx; we will thus have to start from a time-domain representation of the waveform as a train of laryngeal pulses. If an algorithm or device works in a speech production oriented way, it measures individual *laryngeal excitation cycles* or, if some averaging is performed, it determines the *rate of vocal-fold vibration.* The signal-processing point of view can be characterized in such a way that (quasi-)periodicity is observed in the signal, wherever that signal comes from, and that the task is just to extract those features that best represent this periodicity. The pertinent terms are *fundamental frequency* or *fundamental period.* If individual cycles are determined, we may (somewhat inconsistently) speak of *pitch periods* or simply of *periods.* The perception point of view leads to a frequency-domain representation since pitch sensation corresponds to a frequency and not to an average period or a sequence of periods (Goldstein, 1973; Terhardt, 1979; Plomp, 1976). This point of view is associated with the original meaning of the term *pitch.* However, the term pitch has consistently been used as some kind of "common denominator", i.e., as a general name for all those terms mentioned before, at least in the technical literature (Kohler, 1982). In addition, psychoacousticians have started to create new terms for describing the aspects of pitch perception, such as *spectral pitch* or *virtual pitch* (Terhardt, 1979), mostly because they felt it necessary to specify partial aspects of the complex phenomenon of pitch perception more precisely, but also in order to avoid confusions. In the following, we will therefore use the term *pitch* in this wider sense wherever a more restricted description is undesirable or impossible, and take the more precise terms otherwise.

Defining the different representations of pitch, it appears reasonable to proceed from production to perception. Going in that direction we will start at a local and detailed representation and arrive at a more global representation in the case of the perception-oriented view. The basic definitions could thus read as follows (Hess, 1983:475, 1992; Hess and Indefrey, 1987):

> *T_0 is defined as the elapsed time between two successive laryngeal pulses. Measurement starts at a well specified point within the glottal cycle, preferably at the point of*

glottal closure or – if the glottis does not close completely – at the point where the glottal area reaches its minimum. (1)

PDAs that obey this definition will be able to locate the point of glottal closure and to delimit individual laryngeal excitation cycles. This task, which usually forms part of a glottal inverse filter, goes far beyond the scope of ordinary pitch determination; if the speech signal alone is available for the analysis, reliable results are to be expected only for selected algorithms and only if the signal is totally undistorted. With the aid of an instrument this problem can be solved in a more general way.

T_0 is defined as the elapsed time between two successive laryngeal pulses. Measurement starts at an arbitrary point within the glottal cycle. Which point that is depends on the individual method, but for a given PDA this point is always located at the same position within the glottal cycle. (2)

Ordinary time-domain PDAs follow this definition. The reference point can be a significant extreme, a certain zero crossing, an excursion cycle, and so on. This is not necessarily the point of glottal closure itself. Usually, however, it is possible to derive the point of glottal closure from this reference point when the signal is undistorted. Yet the presence of phase distortions can even destroy this possibility. PDAs that follow this definition usually track the signal period by period in a synchronous way, and a commonly used term (although somewhat inconsistent with the definition of the term *pitch* as given above) for what is measured here is *individual pitch periods.*

T_0 is defined as the elapsed time between two successive laryngeal cycles. Measurement starts at an arbitrary instant which is fixed according to external conditions, and ends when a complete cycle has elapsed. (3)

This is an *incremental* definition of T_0. T_0 is still defined as the length of an individual period, but no longer from the speech production point of view, since the definition has nothing to do with an individual excitation cycle. The synchronous way of processing is maintained, but the phase relations between the laryngeal waveform and the markers, i.e., the pitch period delimiters at the output of the algorithm are lost. Once a reference point in time has been established, it will be kept only as long as the measurement is correct and as long as voicing continues. If there is a measurement error, or if voicing ceases, the location of the reference point is lost, and the next reference point may be completely different with respect to its position within the excitation cycle.

T_0 is defined as the average length of several periods, i.e., as the average elapsed time between a small number of successive laryngeal cycles. In which way the averaging is performed, and how many periods are involved, is a matter of the individual algorithm. (4a)

This is the standard definition of T_0 for any PDA that applies stationary short-term analysis, including the implementations of frequency-domain PDAs. Well-known methods, such as cepstrum (Noll, 1967) or autocorrelation (Rabiner, 1977) follow this definition. The corresponding frequency-domain definition reads as follows.

F_0 is defined as the fundamental frequency of an (approximately) harmonic pattern in the (short-term) spectral representation of the signal. It depends on the particular

method whether F_0 is calculated as the frequency of a certain harmonic divided by the respective harmonic number m (including m=1), as the frequency difference between adjacent spectral peaks, or as the greatest common divisor of the frequencies of the individual harmonics. (4b)

The perception point of view of the problem leads to a different definition of pitch. Pitch perception happens in the frequency domain. According to the existing theories (Plomp, 1976),

F_0 is defined as the frequency of the sinusoid that evokes the same perceived pitch (residue pitch, virtual pitch, etc.) as the complex sound which represents the input speech signal. (5)

This definition is principally different from the previous ones. Above all, it is a *long-term* definition (Terhardt et al., 1982). The pitch perception theories were developed for stationary complex sounds and were only extended toward short pulse trains with varying amplitude patterns and constant frequencies, but not toward signals with varying fundamental frequency. Except for some investigations which indicate that the difference limen for F_0 changes goes up by at least an order of magnitude when time-variant stimuli are involved (Harris and Umeda, 1987; 't Hart, 1981), the question of the behavior of the human ear with respect to *short-term* pitch perception is only partially answered, and our knowledge about what kind of short-term "analysis" is executed in the human ear and how it is executed is still incomplete. Hence even such PDAs that claim to be perception-oriented (e.g., Duifhuis et al., 1982, Hermes, 1988) enter the frequency domain in a similar way as in definition (4b), i.e., by a standard short-term transformation such as the discrete Fourier transform (DFT) with previous windowing of the signal.

Since the results of individual algorithms may be different according to the definition they follow, and since the definitions (1) through (5) are partly given in the time (or lag) domain, partly in the frequency domain, it is necessary to reestablish the relation between the time- and frequency-domain representations of pitch,

$$F_0 = 1 \,/\, T_0 \,, \tag{6}$$

in such a way that when a measurement is carried out in one of the domains, however T_0 or F_0 are defined there, the representation in the other domain will always be established by this equation.

2. Categorizing the Various Principles

We subdivide a PDA into three steps of processing: a) the preprocessor, b) the basic extractor, and c) the postprocessor (McKinney, 1965; Hess, 1983:152). The basic extractor performs the main task: it converts the input signal into a series of pitch estimates. The task of the preprocessor is data reduction and enhancement in order to facilitate the operation of the basic extractor. The postprocessor operates in a more application-oriented way. Some of its typical tasks are error correction, smoothing the pitch contour, or graphic display.

The existing PDA principles can be split up into two gross categories when the input signal of the basic extractor is taken as a criterion. If this signal has the same time base as

the original speech signal, the PDA operates in the time domain. It will thus measure T_0 according to one of the definitions (1) through (3). In all other cases, somewhere in the preprocessor the time domain is left. Since the speech signal is time variant, this cannot be done other than by a short-term transformation; in this case we will usually determine T_0 or F_0 according to definitions (4a,b) or (5); in some rare cases (for instance, AMDF) definition (3) may apply as well. Accordingly, we have the two PDA categories: a) *time-domain* PDAs, and b) *short-term analysis* PDAs.

3. A Brief Look at Short-Term Analysis PDAs

3.1 Principle of Short-Term Analysis and a Categorization of PDAs

In any short-term analysis PDA a *short-term* (or *short-time*) *transformation* is performed in the preprocessor step. The speech signal is split up into a series of frames; an individual frame is obtained by taking a limited number of consecutive samples of the signal $x(n)$ from the starting point, $n=q-K+1$, to the ending point, $n=q$. The frame length, K, is chosen short enough so that the parameter(s) to be measured can be assumed approximately constant within the frame. On the other hand, K must be large enough to guarantee that the parameter remains measurable. For most short-term analysis PDAs a frame thus requires two or three complete periods at least. In extreme cases, when F_0 changes abruptly, or when the signal is irregular, the contradiction of these two conditions can be a source of error (Fujisaki et al., 1986). The frame interval Q, i.e., the distance between consecutive frames (or its reciprocal, the frame rate), is determined in such a way that any significant parameter change is documented in the measurements.

The short-term transformation, so to speak, is intended to behave like a concave mirror which focuses all the scattered information on pitch, as far as it is available within the frame, into one single peak in the spectral domain. This peak is then determined by a peak detector (as the usual implementation of the basic extractor in this type of PDAs). Hence the output signal of the basic extractor is a sequence of average pitch estimates. The short-term transform causes the phase relations between the spectral domain and the original signal to be lost. At the same time, however, the algorithm loses much of its sensitivity to phase distortions and signal degradation. Unfortunately the increased reliability of the algorithm is accompanied by an increased computing effort (which is at least one order of magnitude higher than for a time-domain PDA). Much of this effort goes into the numeric calculation of the transform. Besides the search for reliability, the search for a fast implementation has therefore been an important issue in the design of short-term analysis PDAs.

Not all the known spectral transforms show the desired focusing effect. Those ones which do are in some way related to the power spectrum: correlation techniques, frequency-domain techniques, and a least-squares approach (Fig. 4). Among the correlation techniques we find the well known autocorrelation function which became successful in pitch determination of band-limited signals when it was combined with time-domain center clipping (Sondhi, 1968; Rabiner, 1977). Its counterpart is given by applying a distance func-

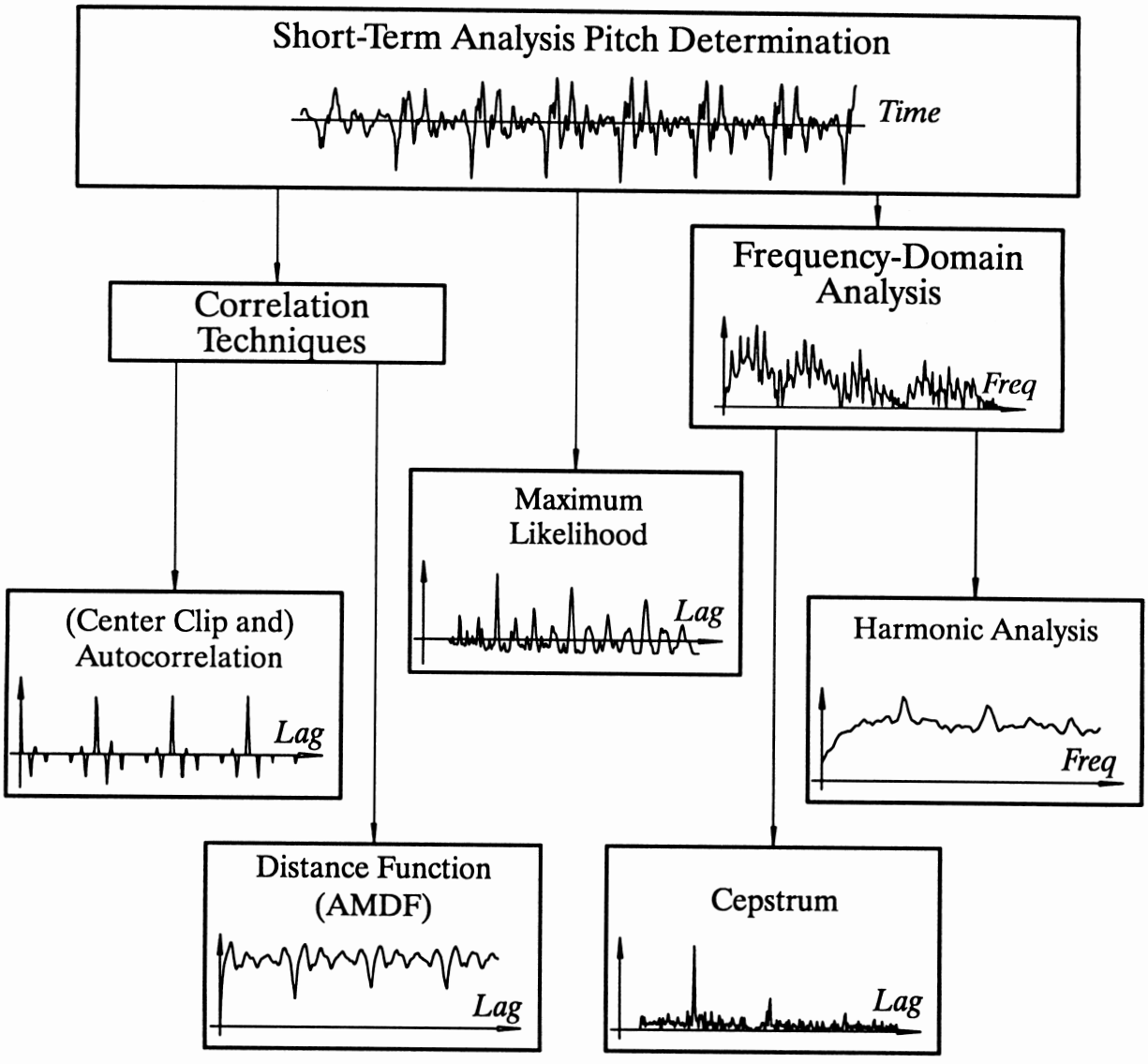


Fig. 4. Methods of short-term analysis (short-time analysis) pitch determination. [Time and lag scales are identical; the frequency scale in the box *Harmonic Analysis* was magnified.]

tion, for instance the average magnitude difference function AMDF (Sobolev and Baronin, 1968; Ross et al., 1974):

$$\mathrm{AMDF}(d) = \sum_{n} |x(n) - x(n+d)| \ . \tag{7}$$

If the signal were strictly periodic, the distance function would take on a value of zero at $d=T_0$. For the quasi-periodic speech signal there will be a strong minimum in the AMDF at this value of the lag (delay time) d. In contrast to all other short-term PDAs where the estimate of T_0 or F_0 is indicated by a maximum whose position *and* value have to be determined, the minimum has an ideal target value of 0 so that we only need to determine its position. For this reason, distance functions do not require (quasi-)stationarity within the measuring interval; they can cope with very short frames of one pitch period or even less.

This principle thus represents the only short-term analysis PDA which is able to follow definition (3) (Moser and Kittel, 1977). The AMDF has also been successfully applied to the linear-prediction residual (Un and Yang, 1977).

The frequency-domain methods are also split up into two groups. Direct determination of F_0 as the location of the lowest peak in the power spectrum is unreliable and inaccurate. It is thus preferred to investigate the harmonic structure of the signal. One way to do this is spectral compression, which computes the fundamental frequency as the greatest common divider of all harmonics. The power spectrum is compressed along the frequency axis by a factor of two, three etc. and then added to the original power spectrum. This operation gives a peak at F_0 resulting from the coherent additive contribution of the higher harmonics (Schroeder, 1968; Noll, 1970; Martin, 1981, 1987). Some of these PDAs stem from theories and functional models of pitch perception in the human ear (Terhardt, 1979; Terhardt et al., 1982; Duifhuis et al., 1982; Hermes, 1988). – The second frequency domain technique leads back into the time domain. Instead of transforming the power spectrum itself (which would lead to the autocorrelation function), however, the inverse transform is performed on the logarithmic power spectrum. This results in the well known *cepstrum* (Noll, 1967), which shows a distinct peak at the "quefrency" (lag) $d=T_0$.

Finally we have to mention the least-squares ("maximum likelihood") approach. This is originally a mathematical procedure to separate a periodic signal of unknown period T_0 (Noll, 1970) from Gaussian noise within a finite signal. Since neither the speech signal is periodic nor the background noise (plus the aperiodic components of the speech signal itself) can be expected as Gaussian, the approach has to be slightly modified in order to work in a PDA (Wise et al., 1976; Friedman, 1977).

In summary, short-term analysis PDAs provide a sequence of average pitch estimates rather than a measurement of individual periods. They are not very sensitive to phase distortions or to absence of the first partial.

3.2 Example: Double-Transform PDA with Nonlinear Distortion in the Frequency Domain

The sensitivity against strong first formants, especially when they coincide with the second or third harmonic, is one of the big problems in pitch determination. This problem is suitable met by some procedure of *spectral flattening.*

Spectral flattening can be achieved in several ways. One of them is time-domain nonlinear distortion, such as center clipping (Sondhi, 1968; Rabiner, 1977). A second way is linear spectral distortion by inverse filtering (Markel, 1972; Un and Yang, 1977). A third way is frequency-domain amplitude compression by nonlinear distortion of the spectrum. This algorithm operates as follows: 1) short-term analysis and transform into the frequency domain via a suitable discrete Fourier transform, 2) nonlinear distortion in the frequency domain, and 3) inverse Fourier transform. The resulting domain is again equivalent to the time domain; to avoid confusion, we will henceforth call it the *lag domain.*

Two members of this group were already mentioned: the autocorrelation PDA (Rabiner, 1977) and the cepstrum PDA (Noll, 1967) which are more closely related than one might conclude from the presentation in Fig. 4. It is well known that the autocorrelation function can be computed as the inverse Fourier transform of the power spectrum. Here,

the distortion consists in taking the squared magnitude of the complex spectrum. The cepstrum, on the other hand, uses the logarithm of the spectrum. The two methods therefore differ only in the characteristics of the respective nonlinear distortions applied in the spectral domain. The cepstrum PDA is known to be rather insensitive to strong formants at higher harmonics but to develop a certain sensitivity with respect to additive noise. The autocorrelation PDA, on the other hand, is insensitive to noise but rather sensitive to strong formants. Regarding the slope of the distortion characteristic, we observe the dynamic range of the spectrum being expanded by squaring the spectrum for the autocorrelation PDA, whereas the spectrum is substantially flattened by taking the logarithm. The two requirements – robustness against strong formants and robustness against additive (white) noise – are in some way contradictory. Expanding the dynamic range of the spectrum emphasizes strong individual components, such as formants, and suppresses wideband noise, whereas spectral flattening equalizes strong components and, at the same time, raises the level of low-energy regions in the spectrum thus raising the level of the noise as well. Thus it is worth while to look for other characteristics that perform spectral amplitude compression. Sreenivas (1981) proposes the 4^{th} root of the power spectrum instead of the logarithm. For larger amplitudes this characteristic behaves very much like the logarithm; for small amplitudes, however, it has the advantage to go to zero and not to $-\infty$. Weiss et al. (1966) use the amplitude spectrum, i.e., the magnitude of the complex spectrum.

Indefrey et al. (1985) implemented these principles together with optional preprocessing to systematically investigate the performance of these PDAs. The four nonlinear spectral functions mentioned before (power spectrum, amplitude spectrum, fourth root of power spectrum, and logarithm) were, among other tests, evaluated using signals with added noise at various noise levels. The PDA was found to break down somewhere between -6 and -12 dB SNR. This value is consistent with data reported elsewhere in the literature for related PDAs (Schroeder, 1968; Noll, 1970; Wise et al., 1976) and shows that there exist a number of short-term PDAs that are extremely noise resistant.

Knowing that many errors arise from a mismatch during short-term analysis (which results in too few or too many pitch periods within a given frame), Fujisaki et al. (1986) investigated the influence of the relations between the error rate, the frame length and the actual value of T_0 for an autocorrelation PDA which operates on the LP residual. The optimum occurs when the frame contains about three pitch periods. Since this value is different for every individual voice, a fixed-frame PDA runs nonoptimally for most situations. For an exponential window, however, this optimum converges to a time constant of about 10 ms for all voices. For a number of PDAs, for example the autocorrelation PDA, such a window permits recursive updating of the autocorrelation function, i.e., even sample-by-sample pitch estimation without excessive computational effort.

4. Time-Domain Pitch Determination Algorithms

This category of PDAs is less homogenous than that of the short-term analysis methods. One possibility to split them up is according to the way how the burden of data reduction

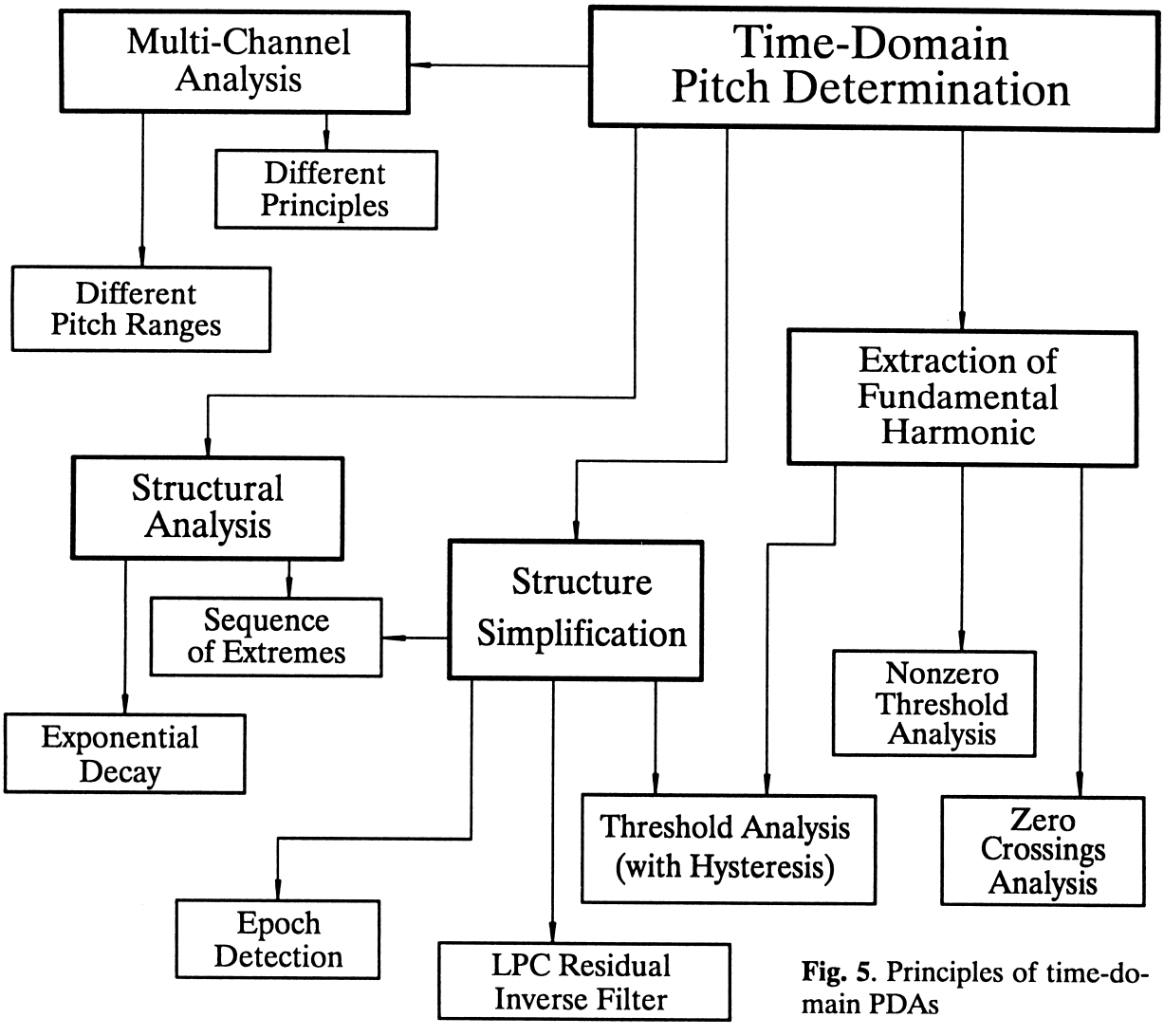


Fig. 5. Principles of time-domain PDAs

is distributed among the preprocessor and the basic extractor. Doing this, we find most time-domain PDAs between two extremes (Fig. 5):

1) The burden is imposed on the preprocessor. In the extreme case, only the waveform of the first harmonic is offered to the basic extractor.

2) The burden is imposed on the basic extractor, which then has to cope with the whole complexity of the temporal signal structure. In the extreme case, the preprocessor is totally omitted.

Time-domain PDAs are principally able to track the signal period by period. At the output of the basic extractor we find a sequence of period boundaries (pitch *markers*). Since the local information on pitch is taken from each period individually, time-domain PDAs are more sensitive to local signal degradations and thus less reliable than the majority of their short-term analysis counterparts. On the other hand, time-domain PDAs may still operate correctly even when the signal itself is irregular due to temporary voice perturbation or laryngealization.

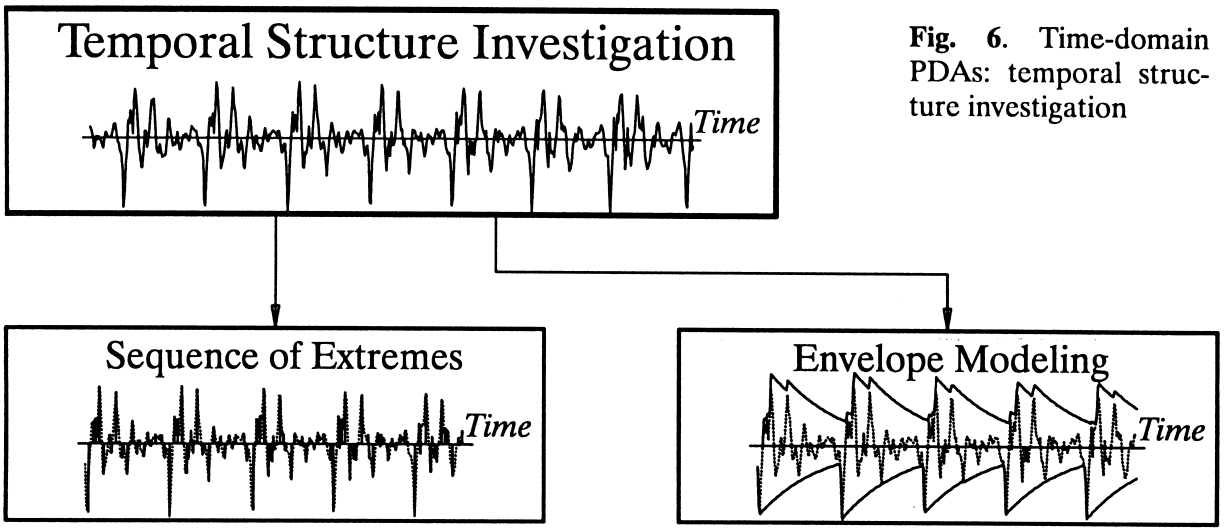


Fig. 6. Time-domain PDAs: temporal structure investigation

4.1 Temporal Structure Investigation

A pitch period is the truncated response of the vocal tract to an individual glottal impulse. Since the vocal tract behaves like a lossy linear system, its impulse response consists of a sum of exponentially damped oscillations. It is therefore to be expected that the magnitude of the significant peaks in the signal is greater at the beginning of the period than versus the end (Fig. 6). Appropriate investigation of the signal peaks (maxima and/or minima) leads to an indication of periodicity.

There are problems associated with this approach, however. First, the frequencies of the dominant damped waveforms are determined by the local formant pattern and may change abruptly. Second, the damping of the formants, particularly of a low first formant, is often quite weak and can be overrun by temporary changes of the signal level. Third, if the signal is phase distorted, different formants may be excited at different points in time. These problems are surmountable, but they lead to relatively complicated algorithmic solutions which have to regard a great variety of temporal structures. Since most of the program instructions are decisions, however, these PDAs run relatively fast. The usual way to carry out the analysis is the following (Reddy, 1967; N.J.Miller, 1975; D. Howard, 1989).

1) Do a moderate low-pass filtering to remove the influence of higher formants.
2) Determine all the local maxima and minima.
3) Exclude those extremes which are found insignificant until one significant point per period is left.
4) Reject obviously incorrect points by local correction.

Structural analysis, especially when many possible structures have to be reviewed, is a good application for self-organizing, nonlinear pattern recognition methods, i.e., for artificial neural networks. Such a PDA was introduced by I. Howard et al. (Howard and Huckvale, 1988, Howard and Walliker, 1989). The speech signal is first divided into 9 subbands with a subsequent half-wave rectification and second-order linear smoothing in each channel. The underlying idea is to obtain a representation similar to that in a wide-band spectro-

gram (cf. Fig. 11). The basic extractor consists of a four-layer perceptron structure, the input layer comprising 41 successive samples with 9 channels each. Two hidden layers with 10 units each and a fully connected network are followed by a one-unit output layer which is intended to yield an impulse when the network encounters a signal structure associated to the instant of glottal closure. The network is trained using (differentiated) output signals of a laryngograph (cf. Sect. 6.2) as reference data. Such a structure has the advantage that it can be based upon several features occurring at different instants during a pitch period. It was shown to outperform conventional devices of the same type, for instance the peak-picking PDA (D. Howard, 1989; see next paragraph) which was evaluated for comparison.

A different solution originates from the analog domain (Dolansky, 1955; Filip, 1969; Winckel, 1964). The envelope of the period is modeled by a cascade of analog differentiators and diode-resistance-capacitance circuits with short rise time constants and comparatively long decay time constants. These circuits emphasize the principal peaks of the signal and suppress all the others. The performance, however, strongly depends on the proper adjustment of the decay time constants. For that reason this relatively simple device works well only for a restricted range of F_0 (about 2 octaves). A manual range switch or something similar is required if a wider range of F_0 is to be analyzed. Due to its simplicity, this principle has been revived in a recent application for cochlear prostheses (D. Howard, 1989). Using a logarithmic amplifier, Howard's PDA avoids a lot of problems associated with the older devices, and his device compares favorably to a number of other PDAs tested for this special application.

4.2 Fundamental Harmonic Processing

F_0 can be detected in the signal via the waveform of the fundamental harmonic. If present in the signal, this harmonic is extracted from the signal by extensive low-pass filtering in the preprocessor. The basic extractor can then be relatively simple. Figure 7 shows the principle of three basic extractors: the zero crossings analysis basic extractor as the simplest device, the nonzero threshold basic extractor, and finally the threshold analysis basic extractor with hysteresis. The zero-crossings analysis basic extractor sets a marker whenever the zero axis is crossed with a defined polarity. This requires that the input waveform has two and only two zero crossings per period. The threshold analysis basic extractor sets a marker whenever a given nonzero threshold is exceeded. The threshold analysis basic extractor with hysteresis acts like the normal threshold analysis basic extractor except that the marker is not set before a second (lower) threshold is crossed in opposite direction. This more elaborate device requires a lesser degree of low-pass filtering in the preprocessor.

The requirement of extensive low-pass filtering is one of two weak points of this otherwise fast and simple principle. For the zero-crossings analysis basic extractor an attenuation of 18 dB/octave is necessary within the range of F_0 to be determined (McKinney, 1965; cf. also Fig. 7). Accordingly, the amplitude of the signal at the basic extractor will vary by more than 50 dB due to the variations of F_0 alone. This dynamic range, increased by the intrinsic dynamic range of the signal (at least another 30 dB), is too much for the PDA to work correctly over the whole range of F_0. The application of a zero-crossings analysis basic extractor thus limits the possible fundamental frequency range. For the threshold analy-

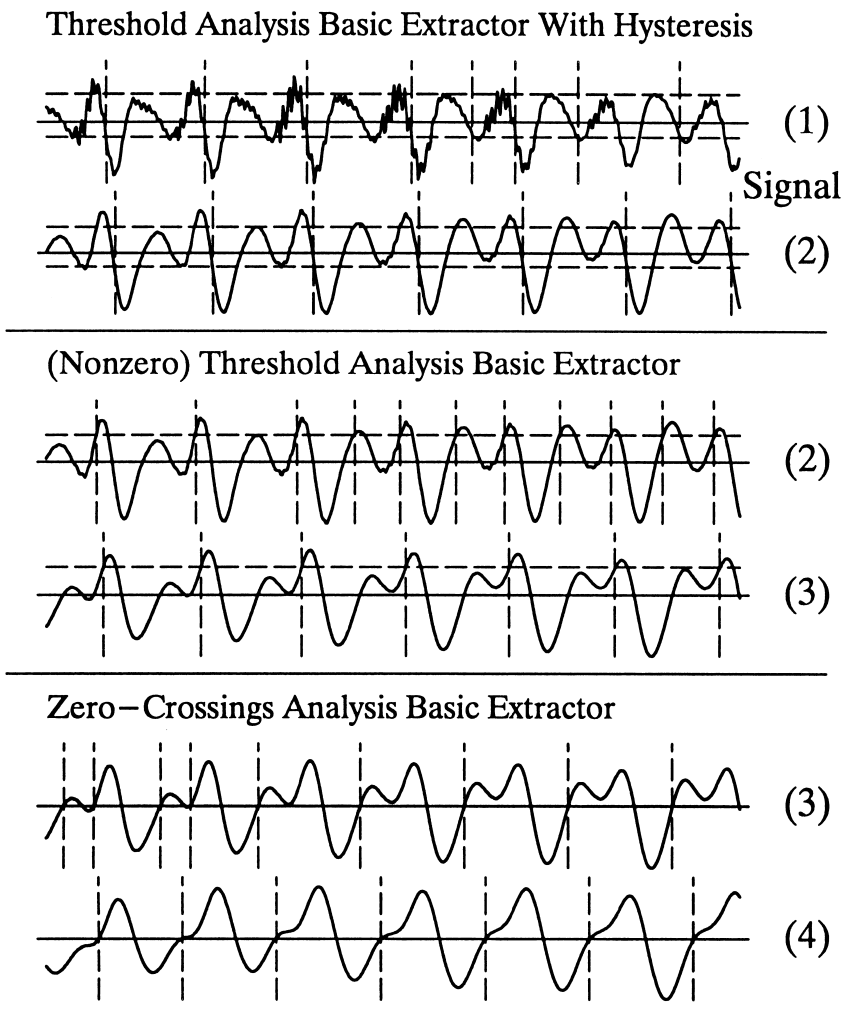


Fig. 7. Example for the performance of basic extractors for fundamental harmonic extraction: zero-crossings analysis basic extractor; nonzero threshold analysis basic extractor; threshold analysis basic extractor with hysteresis. Signals: (1) original, (2) low-pass filtered at 6 dB/octave, (3) low-pass filtered at 12 dB/octave, (4) low-pass filtered at 18 dB/octave. The signal is such that success and failure of the device can be displayed at the same time

sis basic extractor the problem is not so acute, but the fact that the threshold must be adapted to the overall signal amplitude complicates the design of the PDA. In addition there is a systematic measurement artefact associated with the threshold-analysis basic extractor when the amplitude of the input signal varies and the threshold is not properly adapted (Fig. 8). Another inaccuracy (Fig. 9) is intrinsic to the first partial of the signal. When F_0 is close to the formant $F1$, variations of that formant result in time-variant phase distortions of the first partial which will locally change the period duration and with it the T_0 estimate. These inaccuracies are in the order of a few percent; yet they may be intolerable if the respective application requires high accuracy.

In a number of applications, such as voice quality measurement or preparation of reference elements for time-domain speech synthesis (Charpentier and Moulines, 1989), where the signals are expected to be clean, the use of a PDA applying first-partial processing may be advantageous. Dologlou and Carayannis (1989) developed a PDA that overcomes a great deal of the problems associated with the filter necessary to isolate the first partial. An adaptive linear-phase low-pass filter is applied in the preprocessor. This filter consists of a variable-length cascade of second-order filters with a double zero in the z plane at $z=-1$. These filters are consecutively applied to the input signal; after each step the algorithm

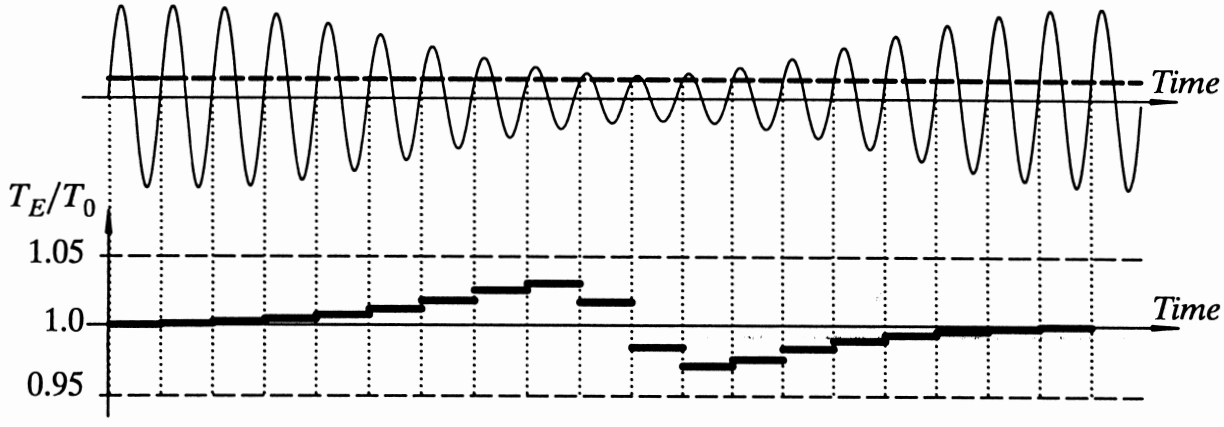


Fig. 8. Measurement errors introduced by changes of the signal amplitude in conjunction with a thr4eshold-analysis basic extractor. Input signal: sinusoid (upper line). (Bottom line) Deviation of the estimate T_E (threshold as in upper line). Only a zero crossings analysis basic extractor or a maximum detector avoids this error

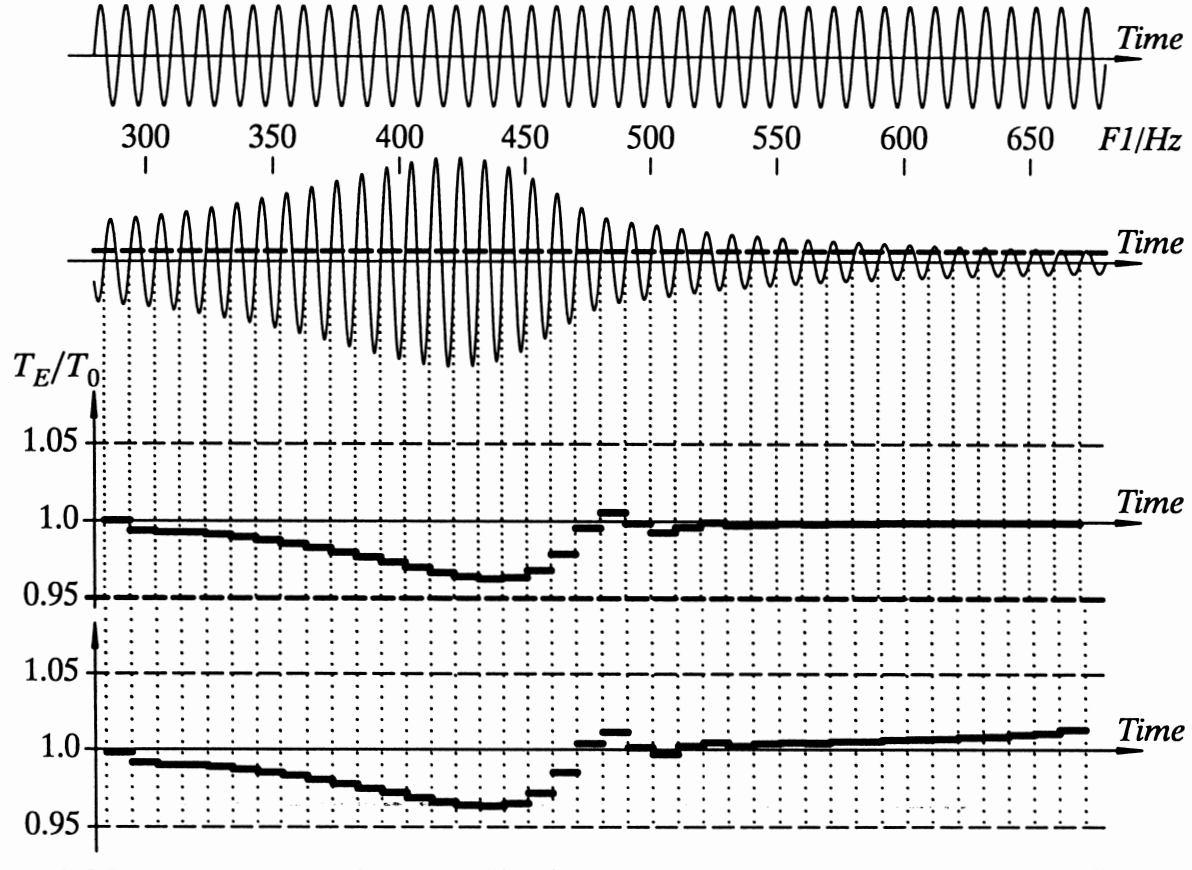


Fig. 9. Measurement errors introduced by changes of the formant $F1$: example. Input signal: sinusoid with the frequency $f=F_0=420$ Hz (upper line). The formant changes from 300 to 650 Hz (this corresponds to a transition [i–a]). The measurement of T_0 is influenced by the phase shift caused by the formant change. (Third line) Deviation of the estimate T_E when a zero-crossings analysis basic extractor is used; (bottom line) same for a threshold-analysis basic extractor (threshold as in second line)

tests whether the higher harmonics are sufficiently attenuated; if yes, the filter stops. T_0 is then derived from the remaining first partial by a simple maximum detector. Very low-frequency noise is tolerable since it barely influences the positions of the maxima.

The second weak point is that this principle is a priori restricted to environments where the first harmonic is present in the signal. There are many applications where this is the case (for instance, in voice quality measurement). If such a PDA, however, is to be applied to processing band-limited signals, the first harmonic must be enhanced or reconstructed. One way to do this is nonlinear distortion. In that respect, many proposals have been made from the beginning on (e.g., Grützmacher and Lottermoser, 1937; Risberg et al., 1960). No single nonlinear characteristic, however, is able to enhance the first harmonic of the signal in an optimal way for any situation, i.e., for any speaker or environmental condition (McKinney, 1965; Hess, 1979); some of them work well in a constrained environment (for instance only with band-limited signals or male voices) or in a realization where several channels with different nonlinear functions are combined (Hess, 1979).

4.3 Simplification of the Temporal Structure

Algorithms of this type take on some intermediate position between the principles of structural analysis and fundamental harmonic extraction. The majority of these algorithms follow one of two principles: a) inverse filtering, and b) epoch detection. Both these principles deal with the fact that the laryngeal excitation function has a temporal structure which is much simpler and more regular than the temporal structure of the speech signal itself, and both methods, when they work, are able to follow definition (1) if the signal is not grossly phase distorted.

The inverse filter approach cancels the transfer function of the vocal tract and thus reconstructs the laryngeal excitation function (cf. also Sect. 5.1). If one is interested in pitch only and not in the excitation function itself, a crude approximation of the inverse filter is sufficient. Such an approximation is realized for instance when the analysis is confined to the first formant (Hess, 1976). The inverse filter approach has one weak point which occurs frequently with female voices. When F_0 is high, it may coincide with the first formant. If the

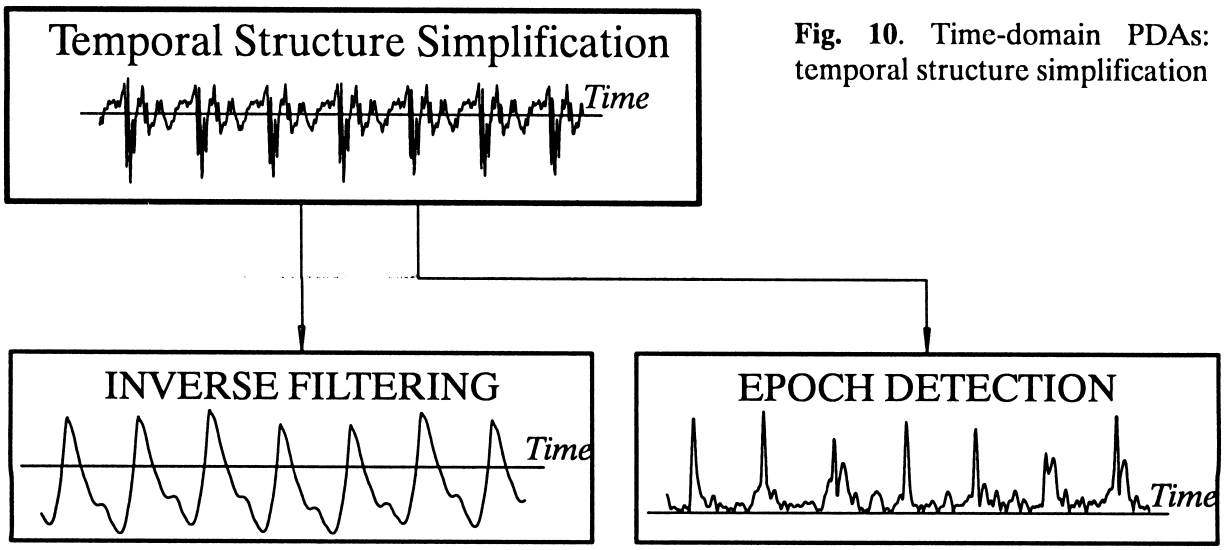


Fig. 10. Time-domain PDAs: temporal structure simplification

inverse filter is not blocked, it then removes the fundamental harmonic (which is extremely strong in this case) from the signal and brings the PDA into failure.

The second principle, epoch extraction, is based upon the fact that at the beginning of each laryngeal pulse there is a discontinuity (in form of an impulse) in the second derivative of the excitation function. Usually this discontinuity cannot be reliably detected in the speech signal due to phase distortions which occur when the waveform passes the vocal tract. The signal is thus first phase shifted by 90° (applying a Hilbert transform). The squares of the original and the phase shifted signals are then added and yield a new signal which represents the instantaneous amplitude of the signal and now shows a distinct peak at the time when the discontinuity in the excitation function occurs. The original method works only when the spectrum of the investigated signal is flat to some extent. To enforce spectral flatness, the analyzed signal is for instance band-limited to high frequencies well above the narrow-band lower formants (Ananthapadmanabha and Yegnanarayana, 1975). Another way is is to analyze the LPC residual (Ananthapadmanabha and Yegnanarayana, 1979) or to filter the signal into subbands (De Mori et al., 1977).

The epoch detection principle depends on the presence of a discontinuity in the second derivative of the laryngeal excitation function. This discontinuity is often weak, especially in back vowels like [u], when a formant exactly coincides with the first or a higher harmonic, or when speech is uttered with a soft or a falsetto voice. In two more recent approaches (Di Francesco and Moulines, 1989; Cheng and O'Shaughnessy, 1989), this drawback was overcome by the finding that the global statistical properties of the waveform change with glottal opening and closing as well. These PDAs, which exploit different features of the signal and were developed independently from each other, derive and apply a generalized maximum-likelihood measure that indicates the instant of glottal closure more precisely than previous epoch-detection PDAs (cf. also Sect. 5.2).

4.4 Multi-channel approaches

Except for the algorithmic investigation of the temporal structure and – nowadays – epoch detection, most simple time-domain PDAs are restricted with respect to the range of F_0 or the type of signal to be processed. One way to increase the range or the reliability of these PDAs is to implement several of them in parallel and to perform some decision as to which one has the "correct" output. The partial PDAs may be identical in design, and each of them may process a subrange of F_0 (McKinney, 1965; Léon and Martin, 1969). On the other hand, they may apply different principles without restriction of the frequency range. The PDAs by Risberg et al. (1960) or by Hess (1979), for instance, use several nonlinear functions to enhance the first harmonic in different ways. Gold and Rabiner (1969) combine several simple peak detection basic extractors together with a pattern-matching procedure. The selection criteria in order to find the most likely channel are defined by a certain channel hierarchy, by a regularity check applying a minimum-frequency selection principle (Risberg et al., 1960; Hess, 1979), by statistical measures (Bruno et al., 1982), or by syntactic rules (De Mori et al., 1977). The selection is continuously checked so that the PDA is able to change its choice at any time.

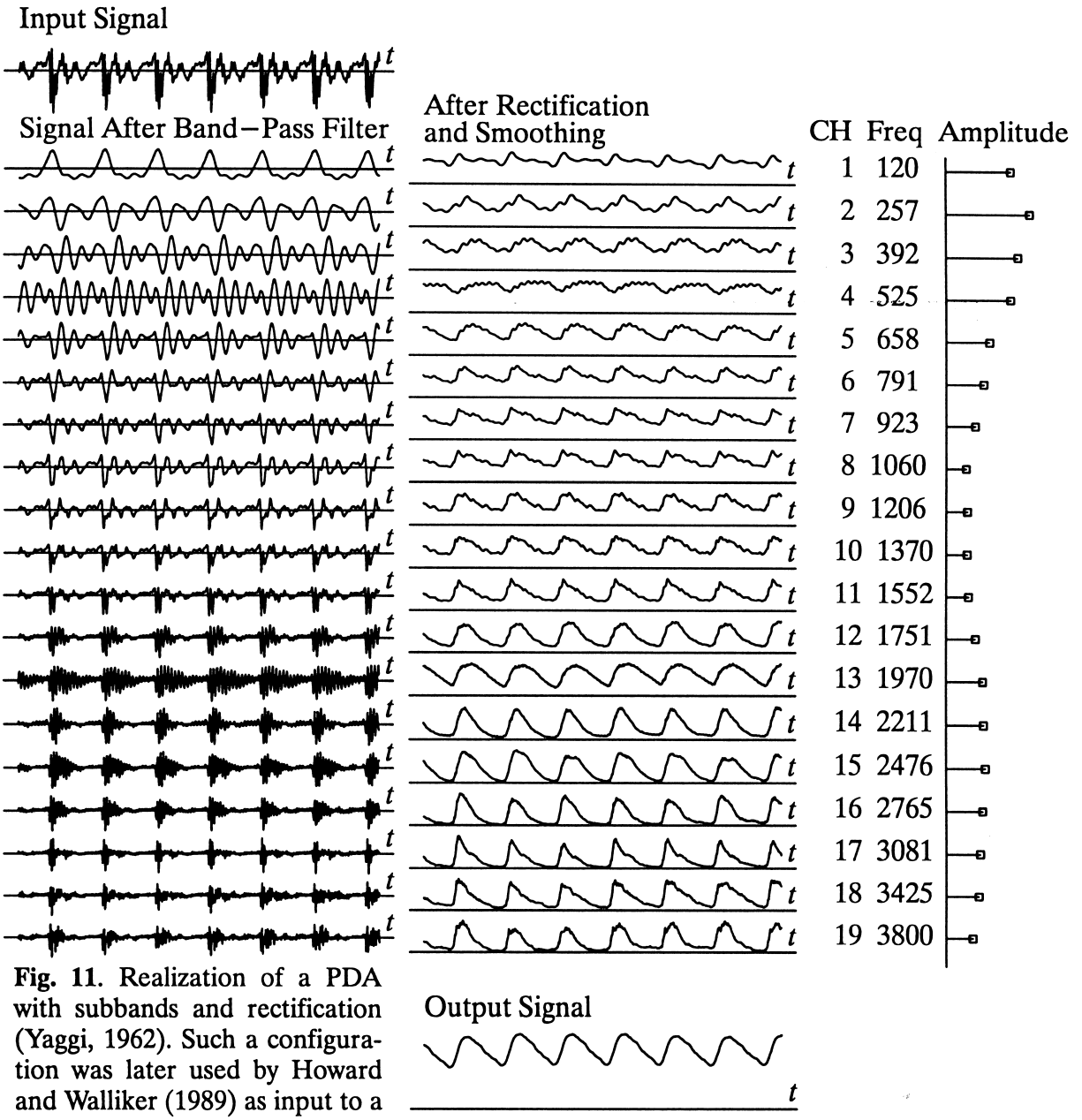


Fig. 11. Realization of a PDA with subbands and rectification (Yaggi, 1962). Such a configuration was later used by Howard and Walliker (1989) as input to a neural network

One problem with time-domain multichannel PDAs is that the individual channels often mark period boundaries at different instants in time, for instance when significant maxima and minima are exploited independently of each other (Gold and Rabiner, 1969). Unless there is a special synchronization routine (Hess, 1979), such PDAs are no longer able to correctly synchronize themselves with the signal and thus have to operate according to the incremental definition (3) or even the short-term definition (4) although they pertain to the time-domain category.

Multi-channel preprocessing by a filter bank dates back to the days of the channel vocoder where the spectral analyzer could also be used as a preprocessor for pitch determination. If the bandwidths of the channels are not too great, there will not be more than

one partial in the lower channels and not more than one formant in the mid and upper channels each. A PDA can thus easily extract the fundamental harmonic once it knows in which channel it is to be found. On the other hand the filter bank output, taken as a whole, behaves in a way similar to a wide-band spectrogram. Those channels which carry the waveforms of the formants coherently reveal maxima of the envelope at the beginning of each pitch period after the instant of glottal closure. This feature can also be exploited for a subsequent structural analysis.

One of the first PDAs of this kind was developed and investigated by Yaggi (1962). Yaggi, however, reported problems with phase distortions in the filter bank. With nowaday's digital filter technology such filter banks can be built as linear-phase networks, and the recent wavelet transform (cf. the PDA by Katambe and Boudreaux-Bartels, 1990), which may be applied like a bank of octave filters, provides another effective means for its implementation. Such a preprocessor (with 9 channels) also serves as the input for the PDA by I. Howard et al. (1989) where the basic extractor is realized by a neural network which performs a structural analysis and is trained to determine the instant of glottal closure.

5. Glottal Inverse Filtering. Determining the Instant of Glottal Closure

5.1 Glottal Inverse Filtering

Glottal inverse filtering is the approximative reconstruction of the excitation signal (the *glottal waveform*) from the speech signal. From the linear model of speech production we know that the voiced speech signal $x(n)$ can be thought of as being generated by the pulse generator characterized by its z transform $P(z)$. The pertinent pulse sequence $p(n)$ passes the glottal shaping filter $G(z)$, at the output of which we have the glottal excitation signal $s(n)$. This signal excites the supraglottal system consisting of the vocal tract $V(z)$ and the radiation component $R(z)$. In terms of transfer functions we obtain

$$X(z) = P(z)\ G(z)\ V(z)\ A\ , \tag{8}$$

where A represents the overall amplitude. A PDA, in this model, can be defined as a device which determines $P(z)$ from $X(z)$. For glottal inverse filtering the task would then read

$$S(z) = P(z)\ G(z) = \frac{X(z)}{V(z)\ R(z)\ A}\ . \tag{9}$$

Thus a filter has to be applied whose transfer function reverts the influence of the vocal tract and the radiation component.

In speech production the radiation component is the low-impedance load which terminates the vocal tract; the volume velocity of the air flow at the lips (and the nose) is converted into sound pressure in the distant field. In a first approximation, which is valid for lower frequencies where the wavelength is large compared to the diameter of the mouth opening, this conversion involves a differentiation, causing a zero at zero frequency. In the inverse filter this zero is reverted by an integrator component, i.e., by a first-order recursive filter with a pole near $z=1$. For reasons of stability, the pole must stay inside the unit circle.

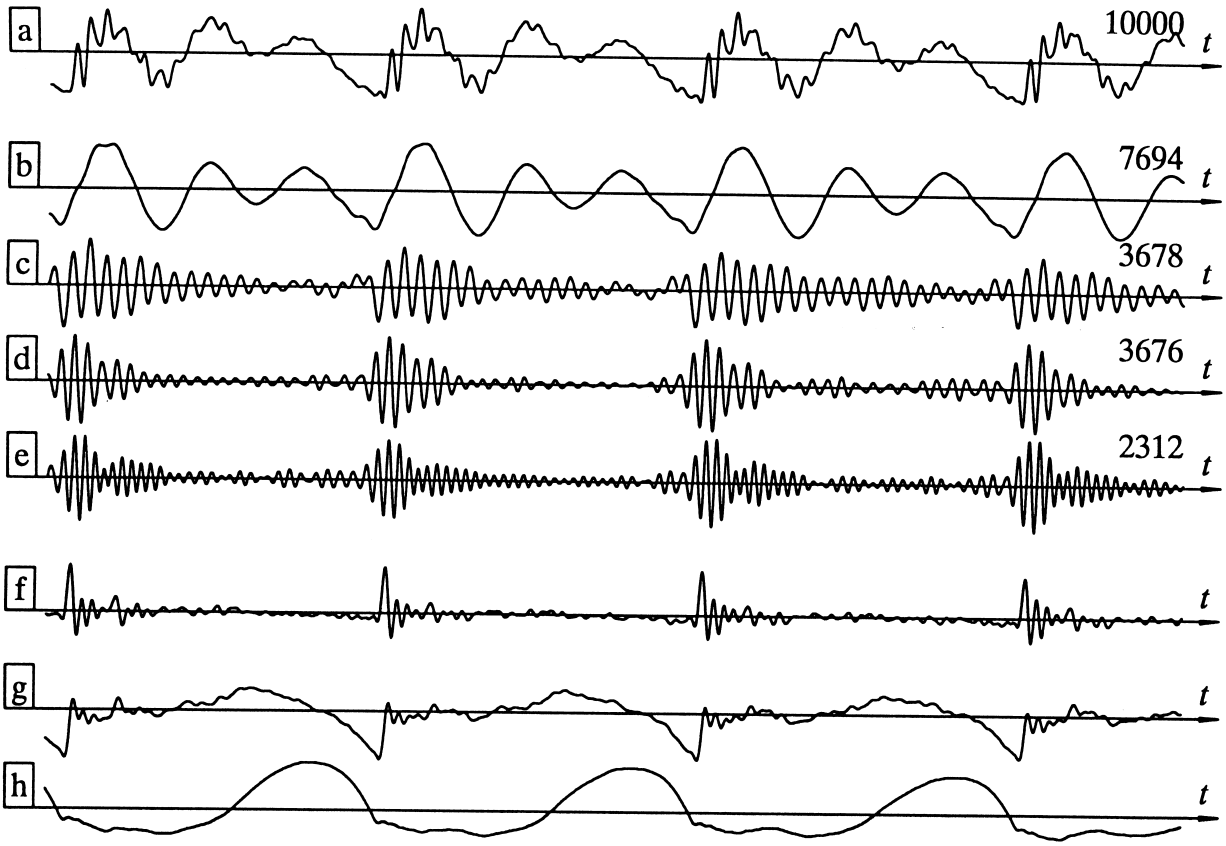


Fig. 12a–h. Inverse–filter analysis. (**a**) Signal: sustained vowel /e/, speaker LJB (male), 32 ms per line; (**b**) waveform of the formant $F1$; (**c–e**) same as (b), this time for the formants $F2$–$F4$; (**f**) differentiated output signal of the inverse filter; (**g**) output signal of the inverse filter; (**h**) reconstructed glottal excitation signal, filtered by the inverse filter and the integrator. The inverse filter was tuned to the following formant frequencies and bandwidths: $F1$=357 Hz, $F2$=2056 Hz, $F3$=2493 Hz, $F4$=3500 Hz; B1=26 Hz, B2=40 Hz, B3=150 Hz, B4=250 Hz. The transfer function of the integrator filter used is $1/H_i(z)=1-0.995z^{-1}$. All signals were normalized before plotting. The numbers on the right–hand side of (a–e) indicate the amplitude of the signal and the individual formants; the amplitude of the signal was normalized to a value of 10000

As glottal inverse filtering is intended to yield a *waveform* rather than certain instants in time, the signal must not at all be phase distorted at low frequencies – a condition nowadays easily met by digital recording equipment.

Glottal inverse filtering requires accurate determination of all formants. For this the following principles have been implemented:

1) individual determination of the different formants, mostly in an interactive way (e.g. Lindqvist, 1965);

2) automatic formant measurement by nonstationary linear-prediction (LPC) analysis during the closed-glottis interval (Wong et al., 1979, Alku, 1992); and

3) cepstrum techniques.

In the classical method (e.g. Lindqvist, 1965), which is carried out in an interactive way, an

antiresonance circuit (i.e., a second-order filter with a complex zero) is provided to cancel each formant individually. The input signal is confined to stationary vowels with significant high-frequency components and formants that are well separable, such as [a] or [ε]. A crude formant analysis provides reasonable initial estimates. Then the antiresonance filters are manually adjusted to the frequencies and bandwidths of the individual formants. Figure 12 shows an example.

A glottal inverse filter using linear-prediction (LPC) analysis was proposed by Wong, Markel, and Gray (1979). Linear prediction models the speech tract as a digital all-pole filter,

$$x(n) = e(n) + \sum_{i=1}^{k} x(n-i) \ , \tag{10}$$

and determines the filter coefficients in such a way that the filter optimally matches the structure of the signal. "Optimally," in this respect, means that the filter has been optimized according to a given criterion. The criterion mostly used involves minimizing the short-term energy of the prediction error, i.e., the energy of the residual signal $e(n)$ within the frame analyzed. This criterion must be further confined for this special application.

Equation (10) says that a sample $x(n)$ can be approximately predicted as the weighted average of the k previous sample of the signal x; $e(n)$ will be the prediction error at the instant n. From the speech production point of view, if $x(n)$ is the speech signal, and if the filter is to serve as a model for the speech tract, then $e(n)$ represents some kind of excitation signal; however, $e(n)$ is usually not identical with the glottal waveform. LPC analysis can be used here when the algorithm is modified in such a way that $e(n)$ represents the glottal waveform itself or at least a waveform having a defined relation to it. The most straightforward way to achieve this is to verify that the LPC filter transfer function $A(z)$ represents the transfer function $V(z)$ of the vocal tract; in this case the residual signal $e(n)$ represents the glottal waveform except for the radiation component, whose reciprocal must be supplied in the form of the first-order integrator filter already known from the earlier discussion in this section.

If $A(z)$ is to represent the vocal-tract transfer function $V(z)$ it is necessary to be certain that the poles of $A(z)$ represent formants only and nothing else. This leads to a modification of the LPC algorithm which involves the following two steps.

1) The poles of $A(z)$ have to be explicitly determined after the analysis; poles that do not pertain to a formant must be excluded from the inverse filter. Routines which perform this task are standard in most scientific program libraries. Once the poles are explicitly known, one can easily assign them to the formants as far as possible and exclude the remainder. One can also exclude a whole frame from further processing if the LPC algorithm has obviously missed a formant (this happens, for instance, when two real poles are supplied instead of a low-frequency or high-frequency formant).

2) In order to represent the vocal-tract transfer function $V(z)$ as accurately as possible, the LPC analysis should be carried out during the closed-glottis interval only. During the open-glottis interval the subglottal system and the vocal tract are coupled via the glottis. This coupling affects the transfer function of the supraglottal system: subglottal formants

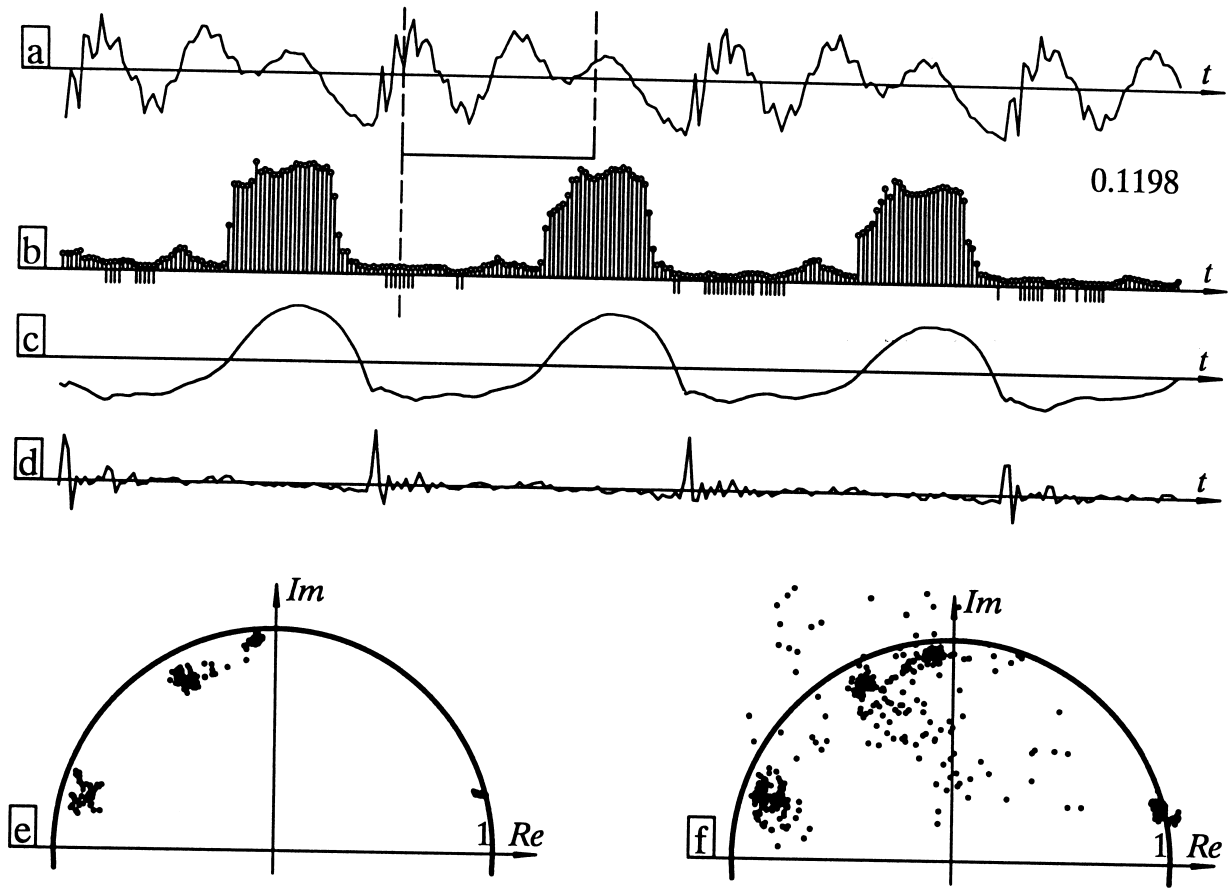


Fig. 13a-f. Glottal inverse filter by Wong et al. (1977, 1979): example of performance. (**a**) Signal: sustained vowel /e/, male speaker, 32 ms per line; same signal as in Fig.12; (**b**) prediction error depending on the starting point q of the frame with the maximum of the normalized error indicated on the right-hand side; (**c**) reconstructed glottal waveform (the integrator being the same as in Fig.12); (**d**) differentiated output signal of the inverse filter; (**e**) locations of the poles of $A(z)$ in the z plane for those cases where $A(z)$ was found appropriate to serve for use in the inverse filter; (**f**) locations of the poles in $A(z)$ for all other cases; (– – – – –) frame selected for computation of the inverse filter. All the frames which pertain to (e) have been marked in (b) by a short continuation line below the baseline. Formant frequencies and bandwidths for the inverse filter applied: $F1$=352 Hz, $F2$=2081 Hz, $F3$=2652 Hz, $F4$=3733 Hz; $B1$=10 Hz, $B2$=109 Hz, $B3$=193 Hz, $B4$=246 Hz. The constraints of the LPC analysis to separate those frames which are suited for selection for the inverse filter (e) and the remainder (f) are rather simple. A frame was excluded from selection when 1) the pertinent LPC filter was not stable, 2) less than 4 formants were detected in the frame, 3) the frame contained a formant frequency below 250 Hz, or 4) one or several formants had excessively large bandwidths. Although there is some variance in the estimates in (e), the formant frequencies and bandwidths are determined rather consistently for the pertinent frames

and antiformants are added to the overall transfer function, and the frequencies and bandwidths of the vocal-tract formants are slightly changed. (Wakita and Fant, 1978). For normal LPC analysis the global estimate is sufficient; here, however, greater accuracy is required.

Compared to an ordinary frame for LPC, the closed-glottis interval is rather short so that the covariance method of linear prediction has to be applied. If the assumption holds that the vocal tract is not excited during the closed-glottis interval, the prediction error will be very low in this case since the vocal tract then represents a linear passive all-pole system. To determine the closed-glottis interval therefore the LPC analysis (using the covariance method and a frame length K which guarantees that $k+K+1$ does not exceed the length of the closed-glottis interval) must be carried out at each sample individually (i.e., using a frame interval equal to the sampling interval of the signal). Low prediction error then indicates that the frame is totally embedded in the closed-glottis interval (cf. Fig. 13).

An alternative criterion for the selection of the closed-glottis interval is the stability of the modeled filter $A(z)$. During the closed-glottis interval the waveforms pertaining to the formants always decay; in this case the LPC filter $A(z)$ will be stable. On the other hand, an instable filter $A(z)$ indicates that there is strong excitation within the analysis interval.

A problem with the algorithm by Wong et al. (1979) is that it requires an LP analysis over the closed-glottis interval. In some voices the closed-glottis interval is very short, or the glottis even does never close completely. This degrades the estimate of the formants and thus the performance of this algorithm. Alku (1992) developed a glottal inverse filter that allows us to perform an iterative LP analysis more globally. First the general slope of the spectrum is approximately flattened by an inverse filter of order 1 to yield an optimal starting point for formant estimation. An LP analysis is carried out over that filtered signal to yield a representation for the transfer function $V(z)$ of the vocal tract. The original signal is then inverse filtered with this filter and passed through an integrator filter. This yields a reasonable estimate of the glottal waveform which is then refined in a second iteration which is almost identical to the first part of the algorithm. It is only now, however, that the frame length is confined to exactly one pitch period ranging from one point of maximal glottal opening (which is determined from the glottal waveform estimate) to the next one. Again the spectrum is flattened using a low-order inverse LP filter, and the vocal-tract transfer function is estimated. Since the algorithm now acts period synchronously, the results are much more accurate than in the first step. Again the original signal is inverse filtered with $1/V(z)$ and passed through an integrator filter to cancel the effect of lip radiation; this yields the final estimate for the glottal waveform.

5.2 Determining the Instant of Glottal Closure

Among all events that characterize the pitch period the *instant of glottal closure* (IGC) occupies a key position. Due to the Bernoulli force exerted on the vocal cords by the air flow in the glottis during the open-glottis interval, the vocal cords are so strongly forced together that they close abruptly and remain closed for about half the glottal cycle (for details see the discussion in Sect.3.1). The air flow is abruptly terminated; this causes a discontinuity in the time derivative of the glottal volume velocity. All formants, particularly the higher ones, are thus simultaneously excited at the IGC. It is thus justified from the speech production point of view to define the *beginning of the pitch period* in the speech signal to coincide with the IGC.

The IGC is rather prominent in normal phonation, i.e., modal register and medium voice effort. It is rather prominent during vocal fry as well. For soft voices as well as for the falsetto register glottal closure still occurs, but somewhat more gradually. In some special cases (breathy voice, certain voice pathologies) the glottis never closes completely. This kind of speech is characterized by weak higher formants. On the other hand, the instant of glottal opening, which passes rather smoothly most of the time, tends to exhibit a second discontinuity (and thus tends to become a second point of excitation) when the voice effort is high.

We can thus expect that the IGC usually represents the most significant and – at the same time – the most easily detectable single event within the pitch period when a reference point with respect to the excitation function is required. In spite of this the task of IGC determination is not at all trivial.

Scanning the PDAs discussed up to now, we see that the algorithms that apply structural simplification (in particular epoch detection) are best suited for IGC determination. In principle most time-domain PDAs place their markers at positions which have some defined relation to the excitation signal. But in many cases this relation is time variant since it depends on the momentary state of the vocal tract. In addition, IGC determination implies the detection of a discontinuity, which is wide-band information, and which is thus masked both by narrow-band formants and high-frequency attenuation in the signal.

The PDA by Ananthapadmanabha and Yegnanarayana (1979) raises the question of the phase of the excitation signal. The ideal case is given when the excitation pulse has a unipolar peak. If the excitation signal is phase shifted by 90°, the IGC coincides with a zero crossing of the excitation pulse, and the amplitude of the pulse is much reduced. This difficulty is overcome by investigating the instantaneous magnitude of the signal which is pulse-like when the spectrum of the signal investigated is approximately flat.

The already-mentioned PDA by I. Howard et al. (1989), which applies a neural network for structural analysis of the output of a filter bank, can be trained toward detecting the IGC. The neural network performs some kind of holistic scan of the structural properties of the signal segment at its input layer and fires at the moment for which it has been trained. This means that the temporal assignment between the temporal structure of the signal and the instant at which the device signals a pitch period boundary is arbitrary and a matter of training. The PDA will thus be trained to detect the IGC when the desired output of the neural net has such a shape that it is close to unity at the IGC and close to zero everywhere else. The differentiated output signal of a laryngograph, after suitable normalization, has this property (cf. Sect. 6.2).

6. Evaluation and Application

To evaluate the performance of a measuring device, one should have another instrument with at least the same accuracy. If this is not available, at least objective criteria – or data – are required to check and adjust the behavior of the new device. In pitch and voicing determination both these bases of comparison are tedious to generate. There is no PDA which operates without errors (Rabiner et al., 1976). There is no reference algorithm, even

with instrumental support, that goes completely without manual inspection or control (Krishnamurthy and Childers, 1986; Hess and Indefrey, 1987). Only rather recently speech databases with reference pitch contours and voicing information have become available (e.g., Carré et al., 1984; Picone et al., 1987), and only then designers of new PDAs started providing detailed data on the performance of their algorithms (e.g. Fujisaki et al., 1986; Indefrey et al., 1985).

6.1 Error Analysis in Pitch Determination

According to the classical study by Rabiner et al. (1976), which established the guidelines for the performance evaluation of these algorithms, PDAs (and voicing determination algorithms, VDAs) commit four types of errors: 1) gross pitch determination errors; 2) fine pitch determination errors, i.e., measurement inaccuracies; 3) voiced-to-unvoiced errors; and 4) unvoiced-to-voiced errors. The latter two types represent errors of voicing determination whereas the first ones refer to pitch determination.

Gross pitch determination errors are "drastic failures of a particular method or algorithm to determine pitch" (Rabiner et al., 1976). Usually an error is regarded to be gross when the deviation between the correct value of T_0 or F_0 and the estimate of the PDA exceeds the maximum rate of change a voice can produce without becoming irregular [Rabiner et al. (1976): 1 ms; Hess and Indefrey (1987): 10%; Krubsack and Niederjohn (1989): 0.25 octave]. On the other hand, errorlike situations may also arise from "drastic failures of the voice to produce a regular excitation pattern," which is not very frequent in well-behaved speech (Dolansky and Tjernlund, 1968), but is nearly always the case when the voice temporarily falls into vocal fry (Fourcin, 1974; Hollien, 1974; Secrest and Doddington, 1982; cf. Fig. 2). Hence, gross errors arise mainly from three standard situations.

1) *Adverse signal conditions:* strong first formants, rapid change of the vocal tract position, band-limited or noisy recordings. Good algorithms reduce these errors to a great extent, but cannot avoid them completely (Rabiner et al., 1976).

2) *Insufficient algorithm performance:* e.g., mismatch of F_0 and frame length (Fujisaki et al., 1986); temporary absence of the key feature in some algorithms.

3) "Errors" that arise from *irregular excitation of voiced signals.* Since most algorithms perform some averaging or regularity check, they can do nothing but fail when the source gets irregular.

When a PDA is equipped with an error detecting routine (and the majority of PDAs are even if no postprocessor is used), and when it detects that an individual estimate may be wrong, it is usually not able to reliably decide whether this situation is a true measurement error – which should be corrected or at least indicated – or a signal irregularity, where the estimate may be correct and should be preserved as it is. This inability of most PDAs to distinguish between the different sources of errorlike situations is one of the great problems in pitch determination yet unsolved.

Measurement inaccuracies cause a noisiness of the obtained T_0 or F_0 contour. They are small deviations from the correct value but can nevertheless be annoying to the listener. Again there are three main causes.

1) *Inaccurate determination of the key feature.* This applies especially to algorithms that exploit the temporal structure of the signal, for instance when the key feature is a principal maximum whose position within a pitch period depends on the formant *F1*.

2) *Intrinsic measurement inaccuracies,* such as the ones introduced by sampling in digital systems.

3) "Errors" from *small fluctuations of the voice* (jitter or shiommer), which contribute to the perception of "naturalness" and should thus be preserved (or even measured).

Voicing errors are misclassifications of the VDA. We have to distinguish between *voiced-to-unvoiced* errors where a frame is classified unvoiced although it is in fact voiced, and *unvoiced-to-voiced* errors with the opposite way of misclassification. This scheme, as established by Rabiner et al. (1976), does not take into account mixed excitation. Voiced-to-unvoiced errors and unvoiced-to-voiced errors must be regarded separately because they are perceptually not equivalent (Viswanathan and Russell, 1984), and the reasons leading to such errors in an actual implementation may be different and even contradictory.

6.2 Developing Reference PDAs with Instrumental Help

A number of former evaluations used a well-known algorithm, for instance the cepstrum PDA, whose performance was known to be good, and compared the algorithm(s) to be tested to the results of that one (Hess, 1983). Rabiner et al. (1976) used an interactive PDA to generate reference data. This procedure proved reliable and accurate but needed a lot of human work. Dal Degan (1982) took the output signal of a vocoder, where the pitch contour was exactly known, as the standard for his PDA evaluation. Bruno et al. (1982) evaluated the performance of a two-channel PDA using the output signal of a mechanic accelerometer which derives the information on pitch from the vibrations of the neck tissue at the larynx. The same device (Stevens et al., 1975) was used by Viswanathan and Russell (1984) for their evaluation of five PDAs. Indefrey et al. (1985) used a laryngograph to yield the signal for generating a reference contour.

Among all the algorithms used for determining a reference pitch contour, those methods appear most efficient which make use of an instrument (such as a mechanic accelerometer or a laryngograph) that derives pitch directly from the laryngeal waveform. This type of algorithm avoids most errors pertinent to the problem of pitch determination from the speech signal, and it permits using natural speech for the evaluation of the performance of PDAs. Among the many instruments available [see (Hess, 1983, Chap. 5) for a survey] the laryngograph (Fourcin and Abberton, 1971; Childers and Krishnamurthy, 1985) is especially well suited for this kind of application. It is robust and reliable, does not prevent the speaker from natural articulation, and gives a good estimate for the instant of glottal closure. A number of PDAs have been designed for this device (e.g. Krishnamurthy and Childers, 1986; Hess and Indefrey, 1987). In addition, Childers et al. (1989) propose a four-category VDA exploiting the speech signal and the laryngogram. In the following, one of these algorithms (Hess and Indefrey, 1987) is presented in some more detail.

The principle of the laryngograph is well known. A small high-frequency electric current is led through the larynx by a pair of electrodes which are pressed against the neck at the position of the larynx from both sides. The opening and closing of the glottis during each

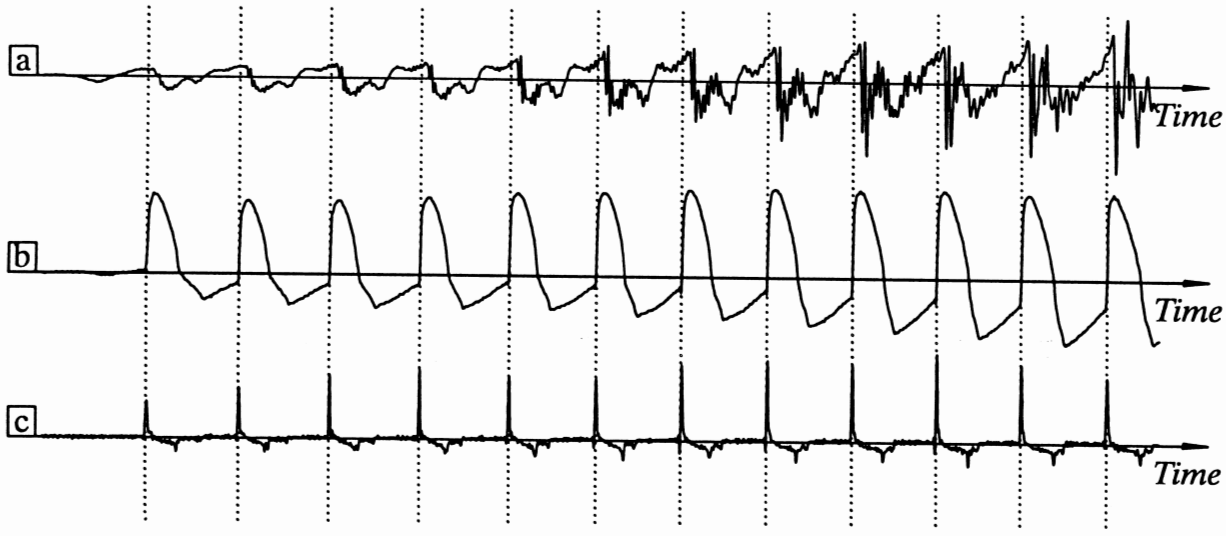


Fig. 14a-c. Speech signal (**a**), laryngogram (**b**), and differentiated laryngogram (**c**). The markers delimiting the individual periods were derived from the maxima of (c). Signal: transition [ja]; speaker WGH (male)

pitch period causes the laryngeal conductance to become time variant; thus the HF current is amplitude modulated. In the receiver the current is demodulated and amplified. Finally, the resulting signal is high-pass filtered in order to remove unwanted low-frequency components due to vertical movement of the larynx

Figure 14 shows an example of the laryngogram (the output signal of the laryngograph) together with the pertinent speech signal. In contrast to the speech signal, the laryngogram is hardly affected by the momentary position of the vocal tract, and the changes in shape or amplitude are comparatively small. Since every glottal cycle is represented by a single pulse, the use of the laryngograph reliably suppresses gross period determination errors. In addition, it supplies the basis for a good voiced-unvoiced discrimination since the laryngogram is almost zero during unvoiced segments where the glottis is always open. Nonetheless, the laryngograph is not free from any problem: it may fail temporarily or permanently for some individual speakers, or it may miss the beginning or end of a voiced segment by a short interval, for instance when the vocal folds, during the silent phase of a plosive, continue to oscillate without producing a signal, or when voicing is resumed after a plosive, and the glottis does not completely close during the first periods (Childers and Krishnamurthy, 1985). For such reason, visual inspection of the reference contour is necessary even with this configuration; these checks, however, can be confined to limited segments of the signal.

What key feature is best used for delimiting the individual periods? According to the theory of voice excitation (van den Berg, 1958; cf. also Stevens, 1977), the instant of glottal closure is the point of maximum vocal-tract excitation, and it is justified to define this instant to be the beginning of a pitch period. In the laryngogram this feature is well documented. As long as the glottis is open, the conductance of the larynx takes on a minimum, and the laryngogram is low and almost flat. When the glottis closes, the laryngeal conduc-

tance goes up, and the laryngogram shows a steep upward slope. The point of inflection during the steep rise of the laryngogram, i.e., the instant of the maximum change of the laryngeal conductance, was found suited best to serve as the reference point for this event.

To press measurement inaccuracies due to signal sampling below the difference limen for perception of F_0 changes over the whole range of F_0, a temporal resolution corresponding to a sampling frequency of more than 100 kHz is required. The strategy of the algorithm is as follows.

1) The laryngogram, originally sampled at 16 kHz, is digitally differentiated by a first-order nonrecursive differentiator filter. The algorithm then determines the significant maxima of the differentiated laryngogram; spurious maxima due to noise are suppressed by simple threshold discrimination.

2) The locations of the maxima of the differentiated laryngogram represent the raw positions of the period delimiters ("markers"); around these points the sampling rate of the signal is increased by a factor of 8; after differentiating and interpolation, the location of the maximum is determined with a temporal resolution of 7.8 μs.

With passband and stopband cutoff frequencies of 5 and 9 kHz, respectively, and a stopband attenuation of more than 72 dB (to keep aliasing distortions below the level of the quantizing noise of the laryngogram), a 144^{th}-order linear-phase interpolator filter proved necessary. The first-order differentiator filter is sufficient to estimate the positions of the raw markers at the original sampling rate of 16 kHz. For the accurate measurement, however, a 7^{th}-order differentiator filter (referring to the original sampling frequency of 16 kHz) gives a good approach to the ideal differentiator in the interesting frequency range below 5 kHz. To minimize roundoff errors, the two filters had to be combined to a 200^{th}-order nonrecursive filter so that differentiation and interpolation finally are performed in one step at the increased sampling frequency of 128 kHz. Same as before, this filter is only applied in the immediate vicinity of the raw markers.

Although the use of the laryngograph reduces the number of gross errors to a minimum, there are still occasional failures of the algorithm in specific situations. Hence a simple error detection logic based on threshold analysis had to be incorporated. First of all, this logic suppresses weak markers that may occur due to noise in the laryngogram. If a rapid vertical movement of the larynx causes a "marker" to be set, this marker will occur in isolation, not embedded in a train of markers as in voiced speech. Hence, if a single marker or a sequence of not more than two markers is detected within an unvoiced interval of at least 200 ms on either side, the logic treats these markers as erroneous and removes them. For the case that this is not yet sufficient, an interactive routine for visual inspection has been provided that may be used to manually accept or reject individual markers that were not reliably accepted or rejected by the automatic procedure.

6.3 Comparative Evaluations – Some More Results

Due to the absence of reliable criteria and systematic guidelines, rather few publications on early PDAs included a quantitative evaluation of the algorithms presented. The main results of the classic study by Rabiner et al. (1976) read as follows.

1) None of the PDAs involved worked without errors, even under good recording conditions. Each PDA had its own "favorite" error; nevertheless, any error condition actually occurred for any of the PDAs.

2) Almost any gross error is perceptible; in addition, unnatural noisiness of a pitch contour is well perceived.

3) The subjective evaluation did not match the preference of the objective evaluation. In fact, none of the objective criteria (number of gross errors, noisiness of the pitch contour, voicing errors) correlated well with the subjective scale of preference.

Hence the question what errors in pitch and voicing determination are the really annoying ones for the human ear remained open. This issue was further pursued by Viswanathan and Russell (1984) who developed objective evaluation methods that are closely correlated to the subjective judgments. The individual error categories are weighted according to the consistency of the error, i.e., the number of consecutive erronoeus frames, the momentary signal energy, the magnitude of the error, and the special context.

Indefrey et al. (1985), concentrating on the evaluation of PDAs only, investigated several short-term PDAs in various configurations. Some of the results were shown further above (Sect. 3.2). In a sequel, Indefrey (1987) added several other PDAs to this evaluation. He showed that in many situations different short-term analysis PDAs behave in a complementary way so that combining them to a multi-channel PDA could lead to a better overall performance.

7. Aspects of Application

The area of speech communication systems is one of the important application areas of pitch determination. Other areas include a) phonetics and linguistics (including musicology), i.e., the measurement of pitch contours as carriers of prosodic, phonetic, and musical information; b) education: training aids for the deaf or teaching aids for foreign languages; and c) the application as a diagnostic aid in voice pathology and phoniatrics. Here determination of source parameters from the signal can serve as a quick and easily accessible help for voice diagnostics and for examining the progress of voice therapy. In phoniatric practice direct measurement and investigation of the speech organs is usual and natural, and pitch determination instruments are a most valuable aid; deriving source parameters from the signal, however, is a hopeful alternative, in particular for early detection of developing voice diseases and for diagnostic evaluation of slight pathologies (Davis, 1978).

Each of these applications has a different profile of requirements (Hess, 1983:521). With respect to these requirements the respective applications can be subdivided according to whether the human ear is the final "customer" of a measured pitch contour or not. If the human ear is at the end of the chain the PDA is a part of, it is crucial to know whether there is a time delay for manual correction permitted or not. There is no time in vocoder systems or in an electronic musical instrument or in the recent application of speech-processing hearing prostheses, e.g., cochlear implants (D. Howard, 1989; Fourcin et al., 1983). There is time for manual correction, on the other hand, in high-quality speech synthesis systems which concatenate original speech data in parametric or waveform-coded repre-

sentation and need accurate pitch determination to manipulate pitch and duration (e.g., Charpentier and Moulines, 1989). Even a laryngograph may be applied for such a purpose (Krishnamurthy and Childers, 1986). In the last few years powerful waveform coding schemes which do not need a PDA at all or only a very rudimentary one have been developed that make a vocoder unnecessary in many applications. Those applications which will continue to require a PDA in speech communication systems, such as hearing prostheses or high-quality speech synthesis from stored data, are more fault tolerant than the vocoder. Future developments in the domain of pitch and voicing determination are thus likely to move away from the search for a new principle that is able "to solve everything" toward improved implementations of known algorithms that are cheap, fast and robust at the same time.

References

Ananthapadmanabha T. V., Yegnanarayana B. (1975): "Epoch extraction of voiced speech." IEEE Trans. ASSP-*23,* 562–569

Ananthapadmanabha T. V., Yegnanarayana B. (1979): "Epoch extraction from linear prediction residual for identification of closed glottis interval." IEEE Trans. ASSP-*27,* 309–319

Alku P. (1992): "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering." Speech Communication *11,* 109–118

Berg, J. van den (1958): "Myoelastic-aerodynamic theory of voice production." J. Speech Hear. Res. *1,* 227–244

Bruno G., Di Benedetto M. G., Gilio A. (1982): "A probabilistic pitch estimation method with combination of several techniques." In *Fortschritte der Akustik, FASE/DAGA'82 Göttingen* (VDE-Verlag, Berlin), 975–978

Carré R., Descout R., Eskénazi M., Mariani J., Rossi M. (1984): "The French language database: defining, planning, and recording a large database." *Proc. IEEE ICASSP-84* (IEEE, New York)

Charpentier F. J. (1986): "Pitch detection using the short-term phase spectrum." *Proc. IEEE ICASSP-86,* paper 3.9 (IEEE, New York)

Charpentier F. J., Moulines E. (1989): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones." in *Proc. EUROSPEECH-89, Paris* (CEP Consultants, Edinburgh, UK), vol. 2, 13–19

Cheng Y. M., O'Shaughnessy D. (1989): "Automatic and reliable estimation of glottal closure instant and period." IEEE Trans. ASSP-*37,* 1805–1815

Childers D. G., Hahn M., Larar J. N. (1989): "Silent and voiced/unvoiced/mixed excitation (four–way) classification of speech." IEEE Trans. ASSP-*37* (Corr.), 1771–1774

Childers D. G., Krishnamurthy A. K. (1985): "A critical review of electroglottography." CRC Critical Reviews in Biomed. Engin. *12,* 131–161

Dal Degan N. (1982): "Vocoder quality: an automatic procedure to measure the performance of pitch extractors." In *Proceedings, Globecom'82*

Davis S. B. (1978): "Acoustic characteristics of normal and pathologic voices." Haskins Labs. Status Reports on Speech Research *54,* 133–164

De Mori R., Laface P., Makhonine V. A., Mezzalama M. (1977): "A syntactic procedure for the recognition of glottal pulses in continuous speech." Pattern Recognition *9,* 181–189

Di Francesco R., Moulines, E. (1989): "Detection of the glottal closure by jumps in the statistical properties of the signal." *Proc. EUROSPEECH-89, Paris,* (CEP Consultants, Edinburgh, UK), 39–42

Dolansky L. O. (1955): "An instantaneous pitch-period indicator." J. Acoust. Soc. Am. *27,* 67–72
Dolansky L. O., Tjernlund P. (1968): "On certain irregularities of voiced speech waveforms." IEEE Trans. AU-*16,* 51–56
Dologlou I., Carayannis G. (1989): "Pitch detection based on zero-phase filtering." Speech Communication *8,* 309–318
Duifhuis H., Willems L. F., Sluyter R. J. (1982): "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception." J. Acoust. Soc. Am. *71,* 1568–1580
Filip M. (1969): "Envelope periodicity detection." J. Acoust. Soc. Am. *45,* 719–732
Flanagan J. L., Saslow M. G. (1958): "Pitch discrimination for synthetic vowels." J. Acoust. Soc. Am. *30,* 435–442
Fourcin A. J., Abberton E. (1971): "First applications of a new laryngograph." Medical and Biological Illustration *21,* 172–182
Fourcin A. J., Douek E., Moore B., Rosen S., Walliker J., Howard D. M., Abberton E., Framton S. (1983): "Speech perception with promontory stimulation." Ann. NY Acad. Sci. *405,* 280–294
Friedman D. H. (1977): "Pseudo-maximum-likelihood speech pitch extraction." IEEE Trans. ASSP-*25,* 213–221
Fujimura O. (1968): "An approximation to voice aperiodicity." IEEE Trans. AU-*16,* 68–72
Fujisaki H., Hirose K., Shimizu K. (1986): "A new system for reliable pitch extraction of speech." *Proc. IEEE ICASSP-86,* paper 34.16 (IEEE, New York)
Gold B. (1977): "Digital Speech Networks." Proc. IEEE *65,* 1636–1658
Gold B., Rabiner L. R. (1969): "Parallel processing techniques for estimating pitch periods of speech in the time domain." J. Acoust. Soc. Am. *46,* 442–448
Goldstein J. L. (1973): "An optimum processor theory for the central formation of the pitch of complex tones." J. Acoust. Soc. Am. *54,* 1496–1516
Grützmacher M., Lottermoser W. (1937): "Über ein Verfahren zur trägheitsfreien Aufzeichnung von Melodiekurven." Akustische Z. *2,* 242–248
Harris M. S., Umeda N. (1987): "Difference limens for fundamental frequency contours in sentences." J. Acoust. Soc. Am. *81,*, 1139–1145
Hart J. 't (1981): "Differential sensitivity to pitch distance, particularly in speech." J. Acoust. Soc. Am. *69,* 811–822
Hermes D. J. (1988): "Measurement of pitch by subharmonic summation." J. Acoust. Soc. Am. *83,* 257–264
Hess W. J. (1976): "A pitch-synchronous digital feature extraction system for phonemic recognition of speech." IEEE Trans. ASSP-*24,* 14–25
Hess W. J. (1979): "Time-domain pitch period extraction of speech signals using three nonlinear digital filters." *Proc. IEEE ICASSP-79,* 773–776 (IEEE, New York)
Hess W. J. (1983): *Pitch determination of speech signals – algorithms and devices* (Springer, Berlin)
Hess W. J. (1992): "Pitch and voicing determination." In *Advances in speech signal processing;* ed. by M. M. Sondhi and S. Furui (Marcel Dekker, New York), 3–48
Hess W. J., Indefrey H. (1984): "Accurate pitch determination of speech signals by means of a laryngograph." *Proc. IEEE ICASSP-84,* paper 18B.1 (IEEE, New York)
Hess W. J., Indefrey H. (1987): "Accurate time-domain pitch determination of speech signals by means of a laryngograph." Speech Commun. *6,* 55–68
Hollien H. (1974): "On vocal registers." J. Phonetics *2,* 125–143
Howard D. M. (1989): "Peak-picking fundamental period estimation for hearing prostheses." J. Acoust. Soc. Am. *86,* 902–910
Howard I. S., Huckvale M. A. (1988): "Speech fundamental period estimation using a trainable pattern classifier." *Proc. Speech-88 (FASE), Edinburgh,* (CEP Consultants, Edinburgh)

Howard I. S., Walliker J. R. (1989): "The implementation of a portable real-time multilayer-perceptron speech fundamental period estimator." *Proc. EUROSPEECH-89, Paris,* (CEP Consultants, Edinburgh, UK), 206–209

Huber D. (1988): *Aspects of the communicative function of voice in text intonation* (PhD Diss., University of Gothenburg, Sweden)

Indefrey H. (1987): *Vergleichende Untersuchungen zur Grundfrequenzbestimmung von Sprachsignalen mit dem Ausgangssignal eines Laryngographen als Referenzgröße.* (Dr.-Ing. Diss.; Tech. University of Munich, Germany)

Indefrey H., Hess W. J., Seeser G. (1985): "Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain." *Proc. IEEE ICASSP-85,* vol. 2, paper 11.12 (IEEE, New York)

Katambe S., Boudreaux-Bartels G. F. (1990): "A pitch detector based on event detection using the dyadic wavelet transform." *Proc. Intern. Conf. on Spoken Language Proc. (ICSLP-90)*, 469–472

Klatt D. (1973): "Discrimination of fundamental frequency contours in synthetic speech: implication for models of speech perception." J. Acoust. Soc. Am. *53,* 8–16

Kohler K. J. (1982): "25 Years of Phonetica." Preface to the special issue on pitch analysis. Phonetica *39* (4–5), 185–187

Krishnamurthy A. K., Childers, D. G. (1986): "Two-channel speech analysis." IEEE Trans. ASSP-*34,*, 730–743

Krubsack D. A., Niederjohn R. J. (1989): "A logarithmic approach to fundamental frequency error measurement in speech." J. Acoust. Soc. Am. *85,* 1782–1784

Léon P., Martin Ph. (1969): *Prolégomènes à l'étude des structures intonatives.* Studia Phonetica #2 (Didier, Paris, Montréal)

Lieberman P. (1963): "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges." J. Acoust. Soc. Am. *35,* 344–353

Lindqvist J. (1965): "Studies of the voice source by inverse filtering." STL–QPSR #2, 8–13 (Royal Inst. of Technol., Stockholm)

Markel J. D. (1972): "The SIFT algorithm for fundamental frequency estimation." IEEE Trans. AU-*20,* 367–377

Martin Ph. (1981): "Détection de F_0 par intercorrélation avec une function peigne." *Journées d'Etude sur la Parole 12,* 221–232 (SFA/GALF, Lannion, France)

Martin Ph. (1987): "A logarithmic spectral comb method for fundamental frequency detection." *Proc. 11th Intern. Congr. on Phonetic Sciences, Tallinn*, paper 59.2 (Estonian Academy of Sciences, Tallinn, Estonia)

McKinney N. P. (1965): *Laryngeal frequency analysis for linguistic research* (Res. Rept. # 14, Communic.Sciences Lab., Univ. of Michigan; Ann Arbor, MI, USA)

Miller N. J. (1975): "Pitch detection by data reduction." IEEE Trans. ASSP-*23,* 72–79

Moser M., Kittel G. (1977): "Rechnergestützte Tonhöhenbestimmung." Folia phoniat. *29,* 119–126

Noll A. M. (1967): "Cepstrum pitch determination." J. Acoust. Soc. Am. *41,* 293–309

Noll A. M. (1970): "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate." In *Symposium on Computer Processing in Communication;* ed. by the Microwave Institute; vol.19, 779–797 (The University of Brooklyn Press, New York)

Picone J., Doddington G. R., Secrest B. G. (1987): "Robust pitch detection in a noisy telephone environment." *Proc. IEEE ICASSP-87,* 1442–1445 (IEEE, New York)

Plomp R. (1976): *Aspects of tone sensation* (Academic Press, London)

Rabiner L. R. (1977): "On the use of autocorrelation analysis for pitch detection." IEEE Trans. ASSP-*25,* 24–33

Rabiner L. R., Cheng M. J., Rosenberg A. E., McGonegal C. A. (1976): "A comparative study of several pitch detection algorithms." IEEE Trans. ASSP-*24,* 399–413

Reddy D. R. (1967): "Pitch period determination of speech sounds." Commun. ACM *10,* 343–348

Risberg A., Möller A., Fujisaki H. (1960): "Voice fundamental frequency tracking." *STL-QPSR* #1, 3–5 (Royal Inst. of Technol., Stockholm)

Ross M.J., Shaffer H.L., Cohen A., Freudberg R., Manley H.J. (1974): "Average magnitude difference function pitch extractor." IEEE Trans. ASSP-*22,* 353–361

Schroeder M. R. (1968): "Period histogram and product spectrum: new methods for fundamental-frequency measurement." J. Acoust. Soc. Am. *43,* 829–834

Secrest B. G., Doddington G. R. (1982): "Postprocessing techniques for voice pitch trackers." *Proc. IEEE ICASSP-82,* 172–175 (IEEE, New York)

Sobolev V. N., Baronin S. P. (1968): "Investigation of the shift method for pitch determination." Elektrosvyaz *12,* 30–36 (in Russian)

Sondhi M. M. (1968): "New methods of pitch extraction." IEEE Trans. AU-*16,* 262–266

Sreenivas T. V. (1981): *Pitch estimation of aperiodic and noisy speech signals.* (Diss., Dept. of Electr. Eng., Indian Inst. of Technology, Bombay)

Stevens K. N. (1977): "Physics of laryngeal behavior and larynx modes." Phonetica *34,* 264–279

Stevens K. N., Kalikow D. N., Willemain T. R. (1975): "A miniature accelerometer for detecting glottal waveforms and nasalization." J. Speech Hear. Res. *18,* 594–599

Terhardt E. (1979): "Calculating virtual pitch." Hearing Research *1,* 155–182

Terhardt E., Stoll G., Seewann M. (1982): "Algorithm for extraction of pitch and pitch salience from complex tonal signals." J. Acoust. Soc. Am. *71,* 679–688

Un C. K., Yang S. C. (1977): "A pitch extraction algorithm based on LPC inverse filtering and AMDF." IEEE Trans. ASSP-*25,* 565–572

Viswanathan V. R., Russell W. H. (1984): Subjective and objective evaluation of pitch extractors for LPC and harmonic-deviations vocoders (BBN Report # 5726, Bolt Beranek and Newman, Cambridge, MA, USA)

Wakita H., Fant G. (1978): "Toward a better vocal-tract model." *STL-QPSR* #1, 9–29 (Royal Inst. of Technol., Stockholm)

Weiss M.R., Vogel R.P., Harris C.M. (1966): "Implementation of a pitch-extractor of the double spectrum analysis type." J. Acoust. Soc. Am. *40,* 657–662

Winckel F. (1964): "Tonhöhenextraktor für Sprache mit Gleichstromanzeige." Phonetica *11,* 248–256

Wise J.D., Caprio J.R., Parks T.W. (1976): "Maximum likelihood pitch estimation." IEEE Trans. ASSP–*24,* 418–423

Wong D. Y., Markel J. D., Gray A. H. (1979): Least-squares glottal; inverse filtering from the acoustic speech waveform." IEEE Trans. ASSP-*27,* 350–355

Yaggi L. A. (1962): *Full duplex digital vocoder* (Texas Instrument, Dallas, TX; 5P14–A62)

Zwicker E., Hess W., Terhardt E. (1967): "Erkennung gesprochener Zahlworte mit Funktionsmodell und Rechenanlage." Kybernetik *3,* 267–272

Cross Correlation and Dynamic Programming for Estimation of Fundamental Frequency

David Talkin
Entropic Research Laboratory, Inc.

February 27, 1995

This paper describes an algorithm for the estimation of voice fundamental frequency (*F0*). While *F0* seems "well defined" for normal voiced speech, in fact there are many situations where the "true" *F0* is ambiguous. Thus, an operational definition of *F0* is called for. The approach described below has been refined by several authors over a period of almost 20 years. Programs embodying various versions of the algorithm have been in use for a comparable length of time and have demonstrated excellent performance.

1 Fundamental Frequency Estimation

Attempts to estimate a voicing state and a fundamental frequency in the speech signal are motivated by a speech production model which views the production mechanism as the concatenation of a quasi-stationary excitation source, and a quasi-stationary linear filter corresponding to the slowly-varying vocal track shape. The model produces unvoiced speech (as in the "s" sound) using white noise as the excitation source, and voiced speech (as in vowels) using a pulse train corresponding to glottal activity. Ideally, the voiced speech signal would be periodic, but in practice, the "true" period is sometimes equivocal, especially if only local-in-time evidence is available. This potential for ambiguity remains regardless of the transformation applied to the signal.

1.1 Normalized Cross Correlation

In the following discussion, we transform the speech signal using the cross-correlation function (*CCF*). Given

$$s_m, \quad m = 0, 1, 2, 3, \ldots,$$

a sampled speech signal with sampling interval T, analysis frame interval t, and a window size w, at each frame we advance $z = t/T$ samples with $n = w/T$

samples in the correlation window. w is chosen to be in the neighborhoos of the expected *F0* period; t is sized to adequately sample the time course of changes in *F0*. The *CCF* of K samples length may then be defined as

$$\phi_{mk} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}}, \quad k = 0, K-1;\ m = iz;\ i = 0, M-1,$$

where

$$e_j = \sum_{l=j}^{j+n-1} s_l^2,$$

and i is the frame index for M frames. Note that $-1.0 \leq \phi \leq 1.0$.

We refer to the value of k as the *lag* and to i as the *frame index*. We can represent ϕ_{mk} graphically by assigning lag to the ordinate, frame index (or time) to the abscissa and the value of ϕ at the corresponding time and lag to the degree of shading, with dark shading representing high values (close to 1.0) and white representing low values (close to -1.0). These graphical representations are referred to as *correllograms*.

An utterance containing clear and problematic voiced speech sections with the corresponding *CCF*'s and correllogram may be seen in Figure 1. The only local evidence for the true *F0* is the location and height of maxima in the *CCF*. A segment of unvoiced speech and its *CCF* may be seen in Figure 1(D). Note that, in general, the *CCF* of voiced speech has maxima with comparable amplitudes at lag intervals corresponding to integer multiples the fundamental period while the *CCF* of unvoiced speech has its most prominent maximum at zero lag. If the *CCF* for the problematic case is viewed in a larger temporal context, as in the correlogram of Figure 1(B), the location of the local maximum corresponding to the "true" *F0* becomes more evident.

Note the following general observations regarding speech and speech *CCF*'s:

1. The local maximum in ϕ corresponding to the "true" *F0* for voiced speech (excepting the maximum at zero lag) is usually the largest and is close to 1.0.

2. When multiple maxima in ϕ exist and have values close to 1.0, the maximum corresponding to the shortest period is usually the correct choice.

3. True ϕ maxima in temporally adjacent analysis frames are usually located at comparable lags, since *F0* is a slowly-varying function of time.

4. The "true" *F0* occasionally changes abruptly by doubling or halving.

5. Voicing tends to change states with low frequency.

6. The largest non-zero-lag maximum in ϕ for unvoiced speech is usually considerably less than 1.0.

7. The short-time spectra of voiced and unvoiced speech frames are usually quite different.

Historically, these characteristics and analogous ones pertaining to other transformations of the speech, such as the autocorrelation and AMDF, have guided the design of many *F0* estimation algorithms (see [2] for a thorough review), but combining the often conflicting evidence to determine the voicing state (voiced or unvoiced) and, if voiced, the *F0*, has been an ongoing problem. In the final analysis, the problem is not completely soluble, since the assumptions of a two-state voicing model and a single *F0* are both oversimplifications. On the other hand, the partial solutions achieved so far have led to practical developments in speech technology and provide measurements useful to those studying the basic properties of human speech and the voice.

1.2 Dynamic Programming

Dynamic programming [7] provides a computational framework for integrating the contextual and local evidence available in the correlogram to arrive at a globally best estimate of voicing state and *F0*. Apparently the first reported use of dynamic programming in a similar context was the DYPTRACK algorithm, based on the AMDF and dynamic programming [1], but this work was not made public at the time. This general approach to parameter estimation is clearly outlined by Ney [4]. The approach to *F0* estimation described below is similar to that first publically described by Secrest and Doddington in 1983 and shown by them to have excellent performance simultaneously estimating *F0* and the voicing state [5, 6].

Dynamic programming may be applied to the *F0* problem as follows: Let:

I_i be the number of states hypothesized at frame i, which is one plus the number of non-zero-lag local maxima in ϕ at frame i. Thus, at each frame, $I_i - 1$ *F0*'s (voiced states) and one unvoiced state will be hypothesized.

C_{ij} be the value of the j^{th} non-zero-lag local maximum in ϕ at frame i.

L_{ij} be the lag at which C_{ij} was observed.

We may now define an objective function as the local cost for hypothesizing that frame i is voiced with period tL_{ij} as

$$d_{ij} = \alpha(1 - C_{ij}) + \beta L_{ij},\ 1 \leq j < I_i,$$

while the cost for the distinguished unvoiced hypothesis at frame i is

$$d_{iI_i} = \gamma + \alpha \max_j(C_{ij})$$

where α and β are positive constants. This implements observations 1, 2 and 6 by favoring C_{ij} close to 1.0 and shorter lags for voiced frames and C_{ij}

close to zero for unvoiced frames. The constant γ permits adjustment of the liklihood of a voiced decision.

The inter-frame *F0* transition cost δ at frame i when hypotheses j and k at the current and previous frames are both voiced is defined as

$$\delta_{ijk} = \min\{\xi, (\eta + |\xi - \ln(2.0)|), (\nu + |\xi - \ln(0.5)|)\},$$

where

$$\xi = \ln \frac{L_{ij}}{L_{i-1k}}, \quad i \leq j < I_i; \; 1 \leq k < I_{i-1}$$

and η and ν are positive constants. This implements observations 3 and 4 by making cost an increasing function of inter-frame frequency change, but allowing octave jumps at some specifiable cost.

Given observation 7 above, assume we have some scalar stationarity function S_i, $0 \leq S_i \leq 1$, which is a decreasing function of the magnitude of the rate of change of the speech signal's spectrum with time, we would expect S_i to have minima at boundaries between voiced and unvoiced speech segments.

A voicing state transition cost to be applied when the voicing states hypothesized for the current and previous frames differ is now defined as

$$\delta_{iI_ik} = \delta_{ijI_{i-1}} = \psi + \lambda S_i, \quad 1 \leq k < I_{i-1}; \; 1 \leq j < I_i,$$

where ψ and λ are positive constants. This implements observations 5 and 7 by imposing a cost for any voicing state transition, but reducing the cost of the transition when the signal spectrum is changing rapidly.

We may now define the optimal objective function for frame i as

$$D_{ij} = d_{ij} + \min_{k \in I_{i-1}} \{D_{i-1k} + \delta_{ijk}\}, \quad 1 \leq j \leq I_i,$$

with the initial conditions

$$D_{0j} = 0, \quad 1 \leq j \leq I_0; \quad I_0 = 2.$$

For each state at each frame we save the "back pointers"

$$q_{ij} = k_{min},$$

where k_{min} at each frame are the indices, k, which minimize D_{ij}, so that the optimal state sequence can be retrieved. Back pointers from each state at frame i may be traced backwards until they converge to a common, globally optimal state at frame $i - l$, where lt is the latency of the decision. In practice, this decision latency for the *F0* estimation problem is rarely greater than 100ms. Thus, it is feasible to implement *F0* estimators using this algorithm that can operate continuously, in real time, with modest delay. Finally, the *F0* estimate for the frame is

$$F0_i = \frac{1}{tL_{ij}},$$

where the values of j are those which result in the minimum value for D in the region of convergence.

1.3 Discussion

Reasonable values for the constants in the algorithm may be determined using hill climbing techniques on a standard speech database where the *F0* and voicing state have been hand marked (or otherwise reliably determined, for instance using electro glottography). Fortunately, the performance of the algorithm is weakly sensitive to the exact parameter values once the general operating region has been found.

This algorithm permits estimation of *F0* on a cycle-by-cycle basis, since t, the frame step size and w, the correlation window size can both be set smaller than the expected fundamental period. This is in contrast to autocorrelation-based approaches, where the autocorrelation window must be several glottal periods long.

A variety of inter-frame spectral distance measures can serve as the basis for the "stationarity" measure S_i. Secrest and Doddington suggest the use of LPC log area ratios [6]. Good results have been obtained with a stationarity measure defined as:

$$S_i = \frac{1.0}{(itakura(i-1, i+1) - 0.8)(0.05 + |\frac{rms_{i+1} - rms_{i-1}}{rms_{i+1} + rms_{i-1} + .001}|)},$$

where i is the index of the current frame; rms_i is signal RMS in frame i; and $itakura(i, j)$ is the Itakura-Saito distortion measure [3] between frames i and j.

The precision of the *F0* estimation can be considerably improved by parabolic interpolation of the *CCF*. If a parabola is fit to the three points comprising the peak in the *CCF*, the peak of the parabola is a good estimate of the "true" peak of the corresponding continuous *CCF*. Thus, instead of using the computationally expensive approach of increasing the rate at which the speech signal is sampled, one can apply interpolation on the few peaks in the *CCF* that are finally identified as *F0* period markers.

It is important that DC and other very low frequency noise components be removed from the signal prior to application of the *CCF*. Otherwise, these can generate very high correlation values in unvoiced and "silent" regions of the signal, incorrectly encouraging a "voiced" decision. A high-pass filter with zero response at 0 Hz and a half-power corner frequency at 80 Hz has been found to be quite effective.

The computational load of the dynamic programming *(DP)* can be reduced by limiting the number of candidates considered at each frame. The *DP* load grows as the square of the number of candidates (states) in each frame. Thus, instead of considering *all* local maxima in the *CCF* as period candidates, only the highest N need be considered, where N is on the order of 10-20. This significantly reduces the load in the unvoiced regions where there are many local maxima, *none* of which will ultimately contribute to a period estimation!

The computational load of the *CCF* may be reduced by performing it in two stages. Note that for a given window duration and frame rate, the cost of com-

puting ϕ grows as the square of the speech sample rate. Thus, initial estimates of the *CCF* peak locations can be made on a sample-rate-reduced version of the speech signal. The peak locations can then be refined by recomputing the *CCF* at the higher sample rates only in the vicinity of the initial peak estimates and for only the most promising peaks.

1.4 Figure Caption

Figure 1

Waveform (A), correllogram (B) and cross correlation functions (C, D, E) based on a female voice saying "Are any sub...". The cross correlation plots C, D and E, which were computed at .83 sec, .5 sec and .68 sec, respectively show correlation values as a function of correlation lag with zero lag at the extreme left in each plot. In C the "true" peak corresponding to F0 is actually lower in amplitude than the peak at twice the true period. In D, the true peak is the highest non-zero lag peak. Note that the non-zero-lag peaks in the correlation function based on unvoiced speech, seen in E, are all considerably lower than the zero-lag peak. The correllogram, B, shows the correlation value plotted as a function of time (horizontal axis) and lag (vertical axis). Correlations close to one are shown in black; minus one in white. When the time context surrounding the problematic correlation function in E is taken into account by examining the correllogram in B, the correct peak choice is obvious.

References

[1] Bauer, W. R. and Blankenship, W. A., "DYPTRACK – A Noise-Tolerant Pitch Tracker," NASL-S-210, 525 (UNCLASSIFIED, D. O. D.), 1974.

[2] Hess, Wolfgang, *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[3] Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," *Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-23, pp. 67-72, February, 1975.

[4] Ney, H., "Dynamic Programming Algorithm for Optimal Estimation of Speech Parameter Contours", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-13, No. 3., pp. 208-214, March/April, 1983.

[5] Secrest, B. G. and Doddington, G. R., "Postprocessing Techniques for Voice Pitch Trackers," *ICASSP - 1982*, pp. 172-175.

[6] Secrest, B. G. and Doddington, G. R., "An integrated pitch tracking algorithm for speech systems," *ICASSP - 1983*, pp. 1352-1355, Boston.

[7] Viterbi, A. J.,"Error Bounds for Convolutional codes and an asymtotically Optimum Decoding Algorithm," *IEEE Trans. Inf. Theor.* IT-13, 1967, pp. 260-269.

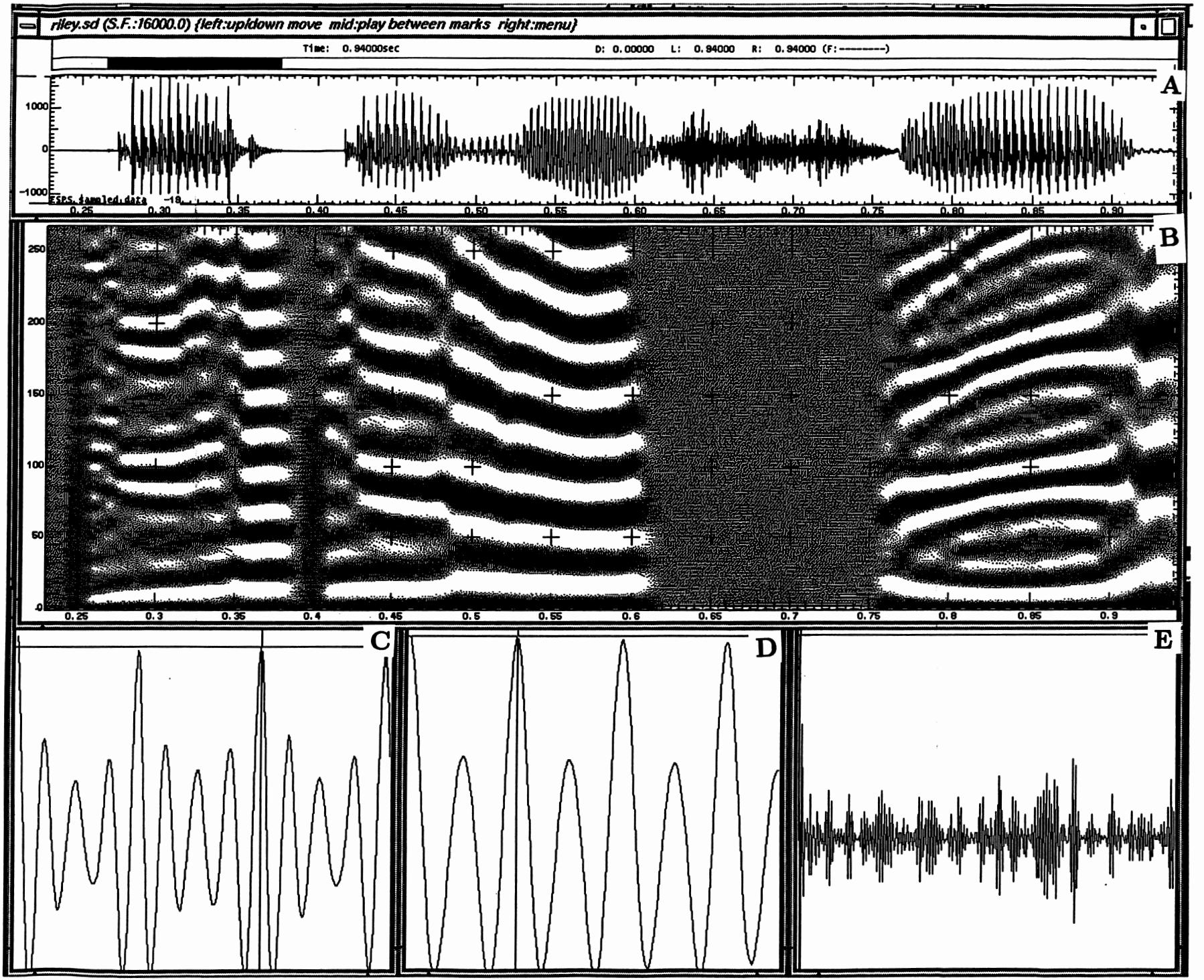

riley.sd (S.F.:16000.0) {left:up/down move mid:play between marks right:menu}
Time: 0.94000sec
D: 0.00000 L: 0.94000 R: 0.94000 (F:--------)
A
B
C
D
E

Figure 1

Rotation-based measure of voice aperiodicity

Paul H. Milenkovic
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
1415 Johnson Drive
Madison, Wisconsin 53706

Abstract

Milenkovic (1987) describes a waveform model for the measurement of the aperiodicity of a voiced speech waveform. The model contains a periodic component, which may vary in amplitude between pitch periods, and a periodicity error (also called the noise component), which has a constant mean-square value across pitch periods. Minimizing the mean-square of the periodicity error provides estimates of 1) pitch period (used to determine jitter), 2) amplitude variation of the periodic component (used to determine shimmer), and 3) magnitude of the aperiodicity noise (used to determine voice SNR).

This report describes how minimizing the periodicity error is equivalent to performing a rotation transformation on signal vectors from two adjoining pitch periods. This transformation is known as SVD in signal processing (Haykin, 1991) and principal components analysis in statistics (Nash, 1979). This connection gives a more numerically stable formula for computing the minimum mean-square error. It also provides a geometric interpretation of the periodic and noise components in relation to the signal vectors, proving the existence of the periodic and noise components in the form required by the model.

1 Introduction

The purpose of this report is to advocate adoption of minimum mean-square error (MSE) waveform matching as a standard for measuring voice aperiodicity. The commercially-available CSpeech software package incorporates a minimum MSE algorithm for determining voice jitter, shimmer, and aperiodicity SNR as described by Milenkovic (1987). Elaboration on the derivation and rationale of this algorithm is warranted. This report also contains numerical recipes to facilitate incorporating minimum MSE into other software packages.

Minimum MSE is an improvement over determining jitter and shimmer by measuring the time and amplitude displacement of a salient waveform peak (Horii, 1979). The time displacement of a zero crossing adjoining a salient peak provides a related way to measure jitter. Besides the difficulty of finding a consistent salient peak from one pitch period to the next, the salient peak measures are sensitive to additive noise in the recording process as well as aperiodicity noise intrinsic to the voice waveform. Performing a minimum MSE waveform match over an entire pulse, however, is a well-known method from radar and sonar engineering for counteracting this noise sensitivity. With voiced speech, this match can take place over an entire pitch period cycle. By employing a sliding pitch period-long analysis frame, it is not necessary to identify the salient peak.

Minimum MSE is also an improvement over conventional approaches to measuring aperiodicity SNR. One method is to measure peaks and valleys of lines in a spectrogram (Muta, *et al*, 1988). The other is to estimate a periodic component by averaging a large number of pitch periods and to measure the noise component as the difference between the speech waveform and the estimated periodic component (Yumoto, *et al.*, 1982). The spectrogram measure requires at least four pitch periods, and the time-domain method requires ten pitch periods. Both methods lump the effects of jitter, shimmer, and aperiodicity noise apart from jitter and shimmer into a generalized measure of noise. Performing a minimum MSE waveform match over one pitch period cycle can reduce the number of pitch periods to two (one cycle matched with an adjoining cycle), and it can help separate the effects of jitter and shimmer from the noise measure.

In employing a minimum MSE waveform match as a unified framework for measuring jitter, shimmer, and aperiodicity SNR, there is some controversy over whether to adopt the seemingly peculiar procedure found in Milenkovic (1987) or to adopt a simpler waveform matching procedure. Such a simpler procedure calls $s(t)$ the waveform in the current pitch period, $s_p(t) = s(t - t_p)$ the waveform in the previous pitch period, and minimizes the error

$$e(t) = s(t) - Ks_p(t) \tag{1}$$

by adjusting amplitude factor K and pitch period t_p (see Qi and Shipp, 1992 for a related method).

The objection to the simpler procedure is how it works when both $s(t)$ and $s_p(t)$ contain aperiodicity noise. The simpler procedure may work for radar where $s_p(t)$ is the noise-free outbound pulse and $s(t)$ is the noisy pulse. When both s and s_p contain noise, the simpler procedure will result in a complicated relation between the minimum mean-square value of e and the true magnitude of the noise contained in both s and s_p. In addition, the relationship between K and waveform shimmer is complicated on account of the bias introduced by the noise.

The more complicated procedure has the advantage that it gives correctly scaled estimates of shimmer and aperiodicity noise if a particular waveform model holds true. This report also shows that the seemingly peculiar algorithm of Milenkovic performs a vector rotation and is therefore identical to the well-known procedure for singular value decomposition (SVD) (Haykin, 1991). SVD is also known as principal components analysis, which has extensive theoretical rationale (Nash, 1979). SVD leads to a geometrical interpretation of signal and noise components, which proves that the model has an exact least-squares instead of only an approximate least mean-square solution as originally supposed.

To widen the application of SVD-based waveform matching, this report purposefully leaves open many other details of a voice analysis system. The initial pitch estimate is such a detail. The waveform matching procedure assumes a rough estimate of the pitch period by other means, and varies t_p to refine the estimate. Another open issue is the question of aligning the analysis frame on pitch epochs. The method of Milenkovic (1987) employs a sliding analysis frame. In a voice analysis system with a reliable means of determining the glottal epoch, the methods described in this report are also applicable to an analysis frame that is aligned on that epoch.

2 Methods

This section of the report 1) describes a waveform periodicity model and reviews the minimum MSE estimate of the model parameters, 2) shows how this model can be reexpressed as a rotation transformation applied to a pair of signal vectors and how the minimum MSE solution can be expressed as the calculation of the optimal rotation that performs SVD, and 3) summarizes this result in the form a numerically-stable recipe for calculation.

2.1 Periodicity model

The waveform $s(t)$ is the speech signal and $s_p(t) = s(t - t_p)$ is the signal from one pitch period before. The quantity t_p is the estimated pitch period, and t_p can be adjusted for a best waveform match between pitch periods. A model of waveform periodicity separates $s(t)$ into a periodic component $p(t)$ and a periodicity error (or noise component) $e(t)$ according to

$$s(t) = p(t) + e(t), \tag{2}$$

$$s_p(t) = p(t - t_p) + e(t - t_p). \tag{3}$$

Furthermore, because the periodic component can vary in amplitude between pitch periods according to $p(t) = Kp(t - t_p)$,

$$s(t) = Kp(t - t_p) + e(t), \tag{4}$$
$$s_p(t) = p(t - t_p) + e_p(t), \tag{5}$$

where $e_p(t) = e(t - t_p)$.

This model is unusual in that the periodic component can have an amplitude modulation K, but then again amplitude modulation is known to be present in speech. If the analysis frame is aligned on a particular glottal epoch, a value of K gets calculated for that alignment. In a sliding analysis frame is used, a new K gets calculated for each updated position.

The waveform matching procedure requires forming vectors of samples

$$\mathbf{s} = [s(\{n_0 - n_p + 1\}T), \ldots, s(n_0T)], \tag{6}$$
$$\mathbf{s}_p = [s_p(\{n_0 - n_p + 1\}T), \ldots, s_p(n_0T)], \tag{7}$$

where T is the interval between waveform samples, n_0 is the integer index controlling position of the analysis frame, and n_p is the number of samples in a pitch period-long frame. In a similar manner, vectors $\mathbf{e}$ and $\mathbf{e}_p$ contain samples of the periodicity error signals $e(t)$ and $e_p(t)$. The vectors $\mathbf{s}$ and $\mathbf{s}_p$ have actual numerical values. The vectors $\mathbf{e}$ and $\mathbf{e}_p$ are only theoretical constructs in the model, but we can estimate their vector magnitudes from observations of $\mathbf{s}$ and $\mathbf{s}_p$.

In the formula $\mathbf{s} - K\mathbf{s}_p$, the periodic component $\mathbf{p}$ (vector of samples of $p(t - t_p)$) simply cancels out, resulting in

$$\mathbf{e} - K\mathbf{e}_p = \mathbf{s} - K\mathbf{s}_p. \tag{8}$$

Next, assume that $\mathbf{e}$ and $\mathbf{e}_p$ are of equal vector magnitude according to $E = \mathbf{e}\mathbf{e}^T = \mathbf{e}_p\mathbf{e}_p^T$ and that they are orthogonal according to $\mathbf{e}\mathbf{e}_p^T = 0$; this is a statement of statistical independence of the noise components in each pitch period. The symbol ^{T} denotes vector transpose and $\mathbf{e}\mathbf{e}^T = \|\mathbf{e}\|^2$ denotes the Cartesian dot product formula for the vector norm square. That $\mathbf{e}$ and $\mathbf{e}_p$ are orthogonal and equal norm permits equating

$$\|\mathbf{e} - K\mathbf{e}_p\|^2 = (1 + K^2)E. \tag{9}$$

That the periodic part cancels out permits equating

$$\begin{aligned} \|\mathbf{e} - K\mathbf{e}_p\|^2 &= \|\mathbf{s} - K\mathbf{s}_p\|^2 \\ &= \mathbf{s}\mathbf{s}^T - 2K\mathbf{s}\mathbf{s}_p^T + K^2\mathbf{s}_p\mathbf{s}_p^T. \end{aligned} \tag{10}$$

Combining these expressions,

$$E = \frac{1}{1+K^2}(\mathbf{s}\mathbf{s}^T - 2K\mathbf{s}\mathbf{s}_p^T + K^2\mathbf{s}_p\mathbf{s}_p^T). \tag{11}$$

We estimate K and t_p and derive noise estimate E by adjusting K and t_p to minimize E. The reason to call this minimum MSE (as opposed to least squares) is the absence of proof that the vectors $\mathbf{p}$, $\mathbf{e}$, and $\mathbf{e}_p$ exist for every given signal vectors $\mathbf{s}$ and $\mathbf{s}_p$. Assuming $e(t)$ and $e_p(t)$ to be independent equal mean-square random processes, and that minimizing E gives an estimate of the minimum MSE, these assumptions provide a weaker criterion for existence of the model. This report will show that the error vectors indeed exist, and that minimizing E gives least-squares error vectors. Even though we cannot uniquely specifiy the error vectors themselves, we can determine their least-squares magnitudes.

The estimation procedure requires stepping through values of t_p, and finding the optimal K for each t_p. In Milenkovic (1987), the procedure is to step through values of t_p that are an integer number of sample intervals T, and to employ parabolic interpolation on the optimal E to find the best t_p between sample positions. An alternative is to generate vector $\mathbf{s}_p$ for the "between" values of t_p by interpolating samples of $s(t)$. In either case, we determine the optimal K for a given t_p by evaluating

$$\frac{\partial E}{\partial K} = \frac{2}{(1+K^2)^2}(K^2\mathbf{s}\mathbf{s}_p^T - K(\mathbf{s}\mathbf{s}^T - \mathbf{s}_p\mathbf{s}_p^T) - \mathbf{s}\mathbf{s}_p^T) = 0. \tag{12}$$

Defining

$$q = \mathbf{s}\mathbf{s}^T - \mathbf{s}_p\mathbf{s}_p^T, \tag{13}$$

$$r = \mathbf{s}\mathbf{s}_p^T, \tag{14}$$

leads to the quadratic

$$K^2 - \frac{q}{r}K - 1 = 0, \tag{15}$$

with solution

$$K = R \pm \sqrt{R^2+1}, \quad R = \frac{q}{2r}. \tag{16}$$

We take the + branch of $\pm$ because that gives a positive value of K, the usual situation with a voiced speech waveform.

This concludes the review of Milenkovic (1987). Next, this solution is reexpressed as a rotation transformation.

2.2 Rotation transformation and SVD

The rotated error vector $\mathbf{e}_r$ is defined as

$$\mathbf{e}_r = \frac{-1}{\sqrt{1+K^2}}(\mathbf{e} - K\mathbf{e}_p) = -s\mathbf{e} + c\mathbf{e}_p \tag{17}$$

for $s = \sin\theta$ and $c = \cos\theta$, where θ is an angle of rotation, and

$$s = \frac{1}{\sqrt{1+K^2}}, \quad c = \frac{K}{\sqrt{1+K^2}}, \tag{18}$$

where $c^2 + s^2 = 1$, the necessary and sufficient condition for s and c to be the sine and cosine of an angle. It also follows that

$$\|\mathbf{e}_r\|^2 = \|\mathbf{e}\|^2 = \|\mathbf{e}_p\|^2 = E. \tag{19}$$

Cancelation of the periodic component $\mathbf{p}$ permits equating

$$\mathbf{e}_r = -s\mathbf{s} + c\mathbf{s}_p. \tag{20}$$

We then estimate s and c by minimizing $\|\mathbf{e}_r\|^2$ subject to the constraint that $c^2 + s^2 = 1$. This is done by simply expressing s and c in terms of the earlier solution for K.

Computing s and c in this manner is identical to determining a two-element principal components analysis. It turns out that expressing s and c in terms of K results in the mathematical formula stated by Nash (1979) for principal components analysis. Expanding

$$\begin{aligned} c^2 = \frac{K^2}{1+K^2} &= \frac{(\sqrt{1+R^2}+R)^2}{2(1+R^2+R\sqrt{1+R^2})} = \frac{(\sqrt{1+R^2}+R)^2}{2\sqrt{1+R^2}(\sqrt{1+R^2}+R)} \\ &= \frac{\sqrt{1+R^2}+R}{2\sqrt{1+R^2}}, \end{aligned} \tag{21}$$

and remembering that $R = \frac{q}{2r}$, it follows from $\sqrt{1+R^2} = \frac{1}{2r}\sqrt{4r^2+q^2}$ that

$$c^2 = \frac{K^2}{1+K^2} = \frac{\sqrt{4r^2+q^2}+q}{2\sqrt{4r^2+q^2}} = \frac{v+q}{2v}, \tag{22}$$

where $v = \sqrt{4r^2+q^2}$.

Taking the positive branch of the square root and setting

$$c = \sqrt{\frac{v+q}{2v}}, \tag{23}$$

the condition $c^2 + s^2 = 1$ requires

$$\begin{aligned} s &= \pm\sqrt{\frac{v-q}{2v}} = \pm\sqrt{\frac{v^2-q^2}{2v(v+q)}} = \pm\sqrt{\frac{v^2-q^2}{4v^2\frac{v+q}{2v}}} \\ &= \pm\sqrt{\frac{r^2}{v^2\frac{v+q}{2v}}} \\ &= \frac{r}{vc}, \end{aligned} \tag{24}$$

where we take the branch of the square root having the same sign as r. This insures the correct result when signal crosscorrelation $r < 0$, a rare occurence with voiced speech that we need to account for anyway.

2.3 Numerical recipe

The numerical algorithm for computing s and c is summarized as follows. Compute

$$r = \mathbf{s}\mathbf{s}_p^T, \tag{25}$$

$$q = \mathbf{s}\mathbf{s}^T - \mathbf{s}_p\mathbf{s}_p^T, \tag{26}$$

$$v = \sqrt{4r^2 + q^2}. \tag{27}$$

The coefficient r is the crosscorrelation between the two pitch periods while q is the signal energy difference between pitch periods.

If $q \geq 0$, the condition where the signal energy is greater than in the previous pitch period, compute

$$c = \sqrt{\frac{v+q}{2v}}; \quad s = \frac{r}{vc}. \tag{28}$$

If $q < 0$, the condition where the signal energy is less than in the previous pitch period, compute

$$s = \mathrm{SGN}(r)\sqrt{\frac{v-q}{2v}}; \quad c = \frac{r}{vs}, \tag{29}$$

where $\mathrm{SGN}(r) = 1$ for $r \geq 0$ (the usual case), $= -1$ for $r < 0$.

The reason for splitting up the solution this way is that it insures computing $v+q$ when $q \geq 0$ (adding two positive numbers) and computing $v - q$ when $q < 0$ (still adding two positive numbers). This avoids the numerical instability resulting from subtracting two (possibly nearly equal) positive numbers.

The formula for E can be reexpressed as

$$E = s^2\mathbf{s}\mathbf{s}^T + c^2\mathbf{s}_p\mathbf{s}_p^T - 2sc\mathbf{s}\mathbf{s}_p^T. \tag{30}$$

In the special case of equal amplitude pitch periods, $\mathbf{s}\mathbf{s}^T = \mathbf{s}_p\mathbf{s}_p^T$ and $c = s = 1/\sqrt{2}$, and the expression simplifies to

$$E = \mathbf{s}\mathbf{s}^T - \mathbf{s}\mathbf{s}_p^T. \tag{31}$$

3 Results

Expressing the minimum MSE solution in terms of principal components analysis leads to a geometric construction. This construction 1) proves the existence of the periodic and noise components, 2) expresses the periodic and noise components in terms of the principal components, and 3) leads to a formal definition of SNR (signal-to-noise ratio) and HNR (harmonics-to-noise ratio).

The principal components are formulated as

$$\begin{bmatrix} \mathbf{s}_r \\ \mathbf{e}_r \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \mathbf{s}_p \end{bmatrix}. \tag{32}$$

The subscript r reminds us that the matrix is a unitary rotation matrix. According to the theory of principal components, when c and s satisfy $c^2 + s^2 = 1$ and $\|\mathbf{e}_r\|^2$ a minimum, $\mathbf{s}_r$ and $\mathbf{e}_r$ are orthogonal principal components. The vectors $\mathbf{s}$, $\mathbf{s}_p$, $\mathbf{s}_r$, and $\mathbf{e}_r$ all lie on an ellipse with $\mathbf{s}_r$, and $\mathbf{e}_r$ marking the major and minor axes.

The existence of the periodicity model is proved by geometric construction. We express the major principle component as

$$\mathbf{s}_r = \mathbf{p}_r + \mathbf{e}_r^{\perp}, \tag{33}$$

where $\mathbf{p}_r$ and $\mathbf{e}_r^{\perp}$ are mutually orthogonal vectors selected from the subspace of vectors orthogonal to $\mathbf{e}_r$ and where $\|\mathbf{e}_r^{\perp}\|^2 = \|\mathbf{e}_r\|^2$. Breaking the major principle component down in this way is possible because $\|\mathbf{s}_r\|^2 \geq \|\mathbf{e}_r\|^2$ by virtue of which component is major (the bigger one) and minor (the smaller one). The magnitudes of $\mathbf{p}_r$ and $\mathbf{e}_r^{\perp}$ are uniquely determined (note that $\|\mathbf{p}_r\|^2 = \|\mathbf{s}_r\|^2 - \|\mathbf{e}_r^{\perp}\|^2 = \|\mathbf{s}_r\|^2 - \|\mathbf{e}_r\|^2$ on account of orthogonality and equality of norms), but the vectors $\mathbf{p}_r$ and $\mathbf{e}_r^{\perp}$ are not unique. That is OK, because we are only interested in the magnitudes for SNR calculations and do not need to recover the actual vectors.

The original vectors are recovered from the principal components according to

$$\begin{aligned} \begin{bmatrix} \mathbf{s} \\ \mathbf{s}_p \end{bmatrix} &= \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \mathbf{s}_r \\ \mathbf{e}_r \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \mathbf{p}_r + \mathbf{e}_r^{\perp} \\ \mathbf{e}_r \end{bmatrix} \\ &= \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \mathbf{p}_r \\ 0 \end{bmatrix} + \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \mathbf{e}_r^{\perp} \\ \mathbf{e}_r \end{bmatrix} = \begin{bmatrix} c\mathbf{p}_r \\ s\mathbf{p}_r \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \mathbf{e}_p \end{bmatrix}. \end{aligned} \tag{34}$$

When the rotation matrix is applied to orthogonal vectors of equal magnitude, the rotation ellipse becomes a circle, and the rotated vectors remain orthogonal with the same magnitude. This allows us to express periodicity error and periodic components of the waveform in terms of principal components. The periodicity error vectors $\mathbf{e}$ and $\mathbf{e}_p$ are orthogonal, equal-magnitude, and rotated versions of minor principal component $\mathbf{e}_r$ and geometrically constructed vector $\mathbf{e}_r^{\perp}$. The major principal component $\mathbf{s}_r$ is the sum of the two constructed vectors $\mathbf{p}_r$ and $\mathbf{e}_r^{\perp}$, and the periodic components are given by $\mathbf{p} = s\mathbf{p}_r$ and $K\mathbf{p} = c\mathbf{p}_r$ for $K = c/s$.

This geometric construction is taking two principle components and generating three vectors: the periodic component (its version scaled by K counts as the same vector) and two independent periodicity error components. As a result of this two-to-three mapping, the construction is not unique, but it exists, the minimum norm of the minor principal component makes the periodicity error components minimum norm, the three vector elements of the periodicity model exist, and the norms (vector magnitudes) are unique.

The proposed definition of periodicity SNR is the average of the energy in the signal for each pitch period divided by the energy in the periodicity error:

$$\begin{aligned}
\text{SNR} &= \frac{1}{2}\frac{\|\mathbf{s}\|^2 + \|\mathbf{s}_p\|^2}{\|\mathbf{e}_r\|^2} \\
&= \frac{1}{2}\frac{(c^2+s^2)\|\mathbf{p}_r\|^2 + 2\|\mathbf{e}_r\|^2}{\|\mathbf{e}_r\|^2} = \frac{1}{2}\frac{\|\mathbf{p}_r\|^2 + 2\|\mathbf{e}_r\|^2}{\|\mathbf{e}_r\|^2} \\
&= \frac{1}{2}\frac{\|\mathbf{s}_r\|^2 + \|\mathbf{e}_r\|^2}{\|\mathbf{e}_r\|^2} \\
&= \frac{1}{2}\left(\frac{\|\mathbf{s}_r\|^2}{\|\mathbf{e}_r\|^2} + 1\right)
\end{aligned} \tag{35}$$

The proposed definition of periodicity harmonics-to-noise ratio (HNR) is the average of the energy in the periodic component for each pitch period divided by the energy in the periodicity error:

$$\begin{aligned}
\text{HNR} &= \frac{1}{2}\frac{(c^2+s^2)\|\mathbf{p}_r\|^2}{\|\mathbf{e}_r\|^2} = \frac{1}{2}\frac{\|\mathbf{p}_r\|^2}{\|\mathbf{e}_r\|^2} \\
&= \frac{1}{2}\frac{\|\mathbf{s}_r\|^2 - \|\mathbf{e}_r\|^2}{\|\mathbf{e}_r\|^2} \\
&= \frac{1}{2}\left(\frac{\|\mathbf{s}_r\|^2}{\|\mathbf{e}_r\|^2} - 1\right)
\end{aligned} \tag{36}$$

4 Conclusions

The waveform matching method of Milenkovic is identical to performing principal components analysis on a pair of signal vectors taken from two adjoining pitch periods. This interpretation provides 1) a numerically stable formula for computation, 2) an interpretation of a voice periodicity model in terms of the principal components, 3) a proof that the method is least squares, 4) a proposed definition for SNR and HNR derived from analysis of two adjoining pitch periods. The proposed definitions and numerical algorithms are applicable to either a sliding analysis frame or an analysis frame aligned with glottal epoch markers.

References

Haykin, S. (1991). Adaptive Filter Theory, Second Edition, Prentice-Hall, Englewood Cliffs, New Jersey.

Horii, Y. (1979). "Fundamental frequency perturbation observed in sustained phonation," J. Speech Hear. Res. 22, 5-19.

Milenkovic, P. (1987). "Least mean square measures of voice perturbation," J. Speech Hear. Res. 30, 529-538.

Muta, H., Baer, T., Wagatsuma, K., Muraoka, T., and Fukuda, H. (1988). "A pitch–synchronous analysis of hoarseness in running speech," J. Acoust. Soc. Am. 84, 1292-1301.

Nash, J. C. (1979). Compact Numerical Methods for Computers, John Wiley, New York.

Qi, Y., and Shipp, T. (1992). An adaptive method for tracking voicing irregularities, J. Acoust. Soc. Am. 91, 3471-3477.

Yumoto, E., Gould, W. J., and Baer, T. (1982). "Harmonics-to-noise ratio as an index of the degree of hoarseness," J. Acoust. Soc. Am. 71, 1544-1550.

SUGGESTION FOR A PITCH EXTRACTION METHOD AND FILE FORMAT FOR PATHOLOGICAL VOICE DATA

Dimitar D. Deliyski

Kay Elemetrics Corp., Dept. of Research and Development
12 Maple Av., Pine Brook NJ 07058, U.S.A.

ABSTRACT

An acoustic model of pathological voice production is presented. It describes the non-linear effects occurring in the acoustic waveform of disordered voices. The noise components such as fundamental frequency and amplitude irregularities and variations, sub-harmonic components, turbulent noise and voice breaks are formally expressed as a result of random time function influences on the excitation function and the glottal filter.

A method for quantitative evaluation of these random functions is described. The method computes some their statistical characteristics which can be useful in assessing voice in clinical practice. More than 33 acoustic parameters are computed, such as: average fundamental frequency, phonatory frequency range, several frequency and amplitude short- and long-term perturbation and variation measures, noise-to-harmonic ratio, voice turbulence and soft phonation indexes, quantitative measures of voice breaks, sub-harmonic components and vocal tremors. This set of parameters, which corresponds to the model, allows a multi-dimensional voice quality assessment. A computer system based on above model and method was developed for the CSL model 4300 (Kay Elemetrics Corp.). A group of 68 people with normal and disordered voices was analyzed using the system in order to define normative values for the acoustic voice parameters.

The file format for voice data used by Kay Elemetrics Corp. is described. This format, which is very similar to a multi-media format supported by Microsoft, allows to keep all the information and associated data in a single file.

1.INTRODUCTION

The classic way to describe the acoustics of human speech is by using the Linear Model of Speech Production [1, 2], where the voice signal is presented as a result of a periodic impulse sequence (excitation) filtered by the glottis, the vocal tract and the lips.

However, the real voice contains irregular components which are (probably) due to the chaotic nature of the laryngeal mechanism [3]. A voice without irregularity is not perceived as human

which is why the advanced speech synthesizers, based on the linear model, introduce some pitch irregularity [4, 14].

2.ACOUSTIC MODEL OF THE PATHOLOGICAL VOICE PRODUCTION

Voice pathology can cause increased noise components in the voice signal such as: fundamental frequency and amplitude irregularities and variations with different patterns, sub-harmonic frequency components, turbulent noise, voice breaks and tremors [2, 5-8]. Understanding the acoustics of these changes is the key to the development of methods for the evaluation of pathologic voices. A formal expression of these changes is given by the *Extended Acoustic Model of the Pathological Voice Production* [9] on Fig.1.

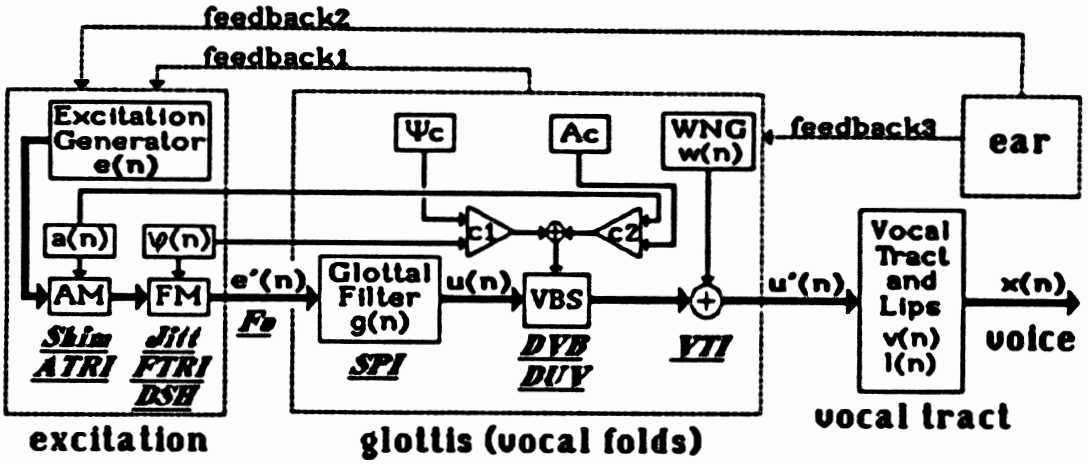


Fig.1:Extended Information Model of the Pathological Voice Production.

The discrete-time formal representation of the model describes the excitation function

$$e'(n) = a(n)\sum_{m=0}^{\infty}\delta[n-\lfloor mTo+\varphi(n)\rfloor]$$

as a modulated impulse sequence, where the frequency modulating (FM) function $\varphi(n)$ and the amplitude modulating (AM) function $a(n)$ are random time functions; $n=0,1...\infty$ is discrete time (samples); To is the period of the sequence (samples); $\delta(n)$ is a Kronecker delta function ($\delta(n=0)=1$, $\delta(n0)=0$); and the carrier sequence is

$$e(n) = \sum_{m=0}^{\infty}\delta(n-mTo).$$

The glottal volume velocity function

$$u'(n) = \begin{cases} u(n)+w(n), if\ a(n) \geq Ac\ and\ \varphi(n) \leq \Psi c \\ w(n), in\ remaining\ cases \end{cases}$$

is a result of filtering of the excitation *e'(n)* by the glottal filter, where

$$u(n) = e'(n) * g(n) = \sum_{m=0}^{\infty} e'(m)g(n-m);$$

$$g(n) = Go(n+1)e^{-cnT}; c \approx 200\pi / \text{sec.};$$

The White Noise Generator (WNG) adds components *w(n)* which model the turbulent components and the Voice Break Switch (VBS) describes the interruptions of the voice generation, where: *g(n)* is the impulse response of the glottal filter, *Go*- scale factor, *T*- sampling period (sec.), *Ac* and *Ψc*- amplitude and frequency break thresholds, c1 and c2- comparators. The convolution of *u'(n)*, the impulse response of the vocal tract filter *v(n)* and the impulse response of the lip-radiation filter *l(n)* results into the modeled voice signal

$$x(n) = u'(n) * v(n) * l(n)$$

where *v(n)* and *l(n)* are considered invariable because it is assumed that the laryngeal pathology does not affect the vocal tract and the lips.

All *a(n)*, *φ(n)* and *w(n)* are random time functions. Therefore the task of acoustic evaluation of pathological voices can be regarded as the extraction of specific statistical parameters of these functions which have clinical significance. The method described below includes three separate parts: pitch extraction (demodulation), noise evaluation and long-term components (tremor) analysis.

3.PITCH EXTRACTION

The amplitude and frequency demodulation curves of the voice signal contain information about the time-domain behavior of *a(n)* and *φ(n)*. The period-to-period pitch extraction [10] is the classic type of demodulation used for evaluation of voice pathology [7, 8]. However the irregularity of the disordered voice makes the pitch extraction inaccurate, often impossible.

In order to provide reliable data an adaptive time-domain pitch-synchronous method for pitch extraction was developed. It consists of the following main steps: fundamental frequency (*Fo*) estimation, *Fo* verification, period-to-period *Fo*-extraction and computation of time-domain voice parameters.

The *Fo-estimation* provides preliminary information about the pitch. It is based on short-term autocorrelation analysis with non-linear *sgn*-coding [11] of the voice signal x(n)

$$R(\tau) = \sum_{n=0}^{N-\tau-1} x'(n)x'(n+\tau),\ 0 \le \tau \le N/2,$$

where: $x'(i)=0$ *if* $Pmin<x(i)<Pmax$;

$x'(i)=1$ *if* $x(i) \ge Pmax$;

$x'(i)=-1$ *if* $x(i) \le Pmin$

and $Pmax=KpAmax$;

$Pmin=KpAmin$;

Amax and *Amin* - global extremes of the current window in the voice signal *x(n)*. The length of the autocorrelation window is 30ms or 10ms depending on the *Fo*-extraction range (67-625Hz or 200-1000Hz). The sampling rate is 50kHz and every window is low-pass filtered at 1800Hz before coding. The value of the coding threshold at this stage of the analysis is $Kp=0.78$ in order to eliminate the incorrect classification of *Fo*-harmonic components as *Fo* [12]. The current window is considered to be voiced with period $To=\tau max$ if the global maximum is $Rmax(\tau max)>KdR(\tau=0)$, where the voiced/ unvoiced threshold value id $Kd=0.27$ [12].

The *Fo-verification* procedure is similar to the *Fo*-estimation. The autocorrelation function is computed again for the same windows at $Kp=0.45$ in order to suppress the influence of components sub-harmonic to *Fo*. The results are compared to the previous step and the decision about the correct *To* is made for all windows where difference is discovered.

A *period-to-period Fo-extraction* is made on the original signal x(n) using a peak-to-peak extraction measurement. It is synchronous with the verified pitch and voiced/unvoiced results computed in the previous steps. A linear 5-point interpolation is applied on the final period-to-period *Fo*-data in order to increase the resolution. This increased resolution is necessary for meaningful frequency perturbation measurements. The peak-to-peak amplitude is also extracted for every period.

The following *time-domain voice parameters* are computed from the extracted pitch data:

Fundamental frequency information measurements: *Average Fundamental Frequency Fo* /Hz/ [2], *Average Pitch Period To* /ms/, *Highest Fundamental Frequency Fhi* /Hz/, *Lowest Fundamental Frequency Flo* /Hz/, *Standard Deviation of the Fundamental Frequency STD* /Hz/ [5], *Phonatory Fundamental Frequency Range PFR* /semi-tones/, *Length of Analyzed Data Sample Tsam* /sec/ and *Number of Pitch Periods PER*.

Short and long-term frequency perturbation functions: *Absolute Jitter Jita* /us/ [13], *Jitter Percent Jitt* /%/ [13], *Relative Average Perturbation RAP* /%/ [7], *Pitch Period Perturbation Quotient PPQ* /%/ [8], *Smoothed Pitch Period Perturbation Quotient sPPQ* /%/ and *Fundamental Frequency Coefficient Variation vFo* /%/ [5].

Short and long-term amplitude perturbation functions: *Shimmer in dB ShdB* /dB/ [13], *Shimmer Percent Shim* /%/ [13], *Amplitude Perturbation Quotient APQ* /%/ [8], *Smoothed Amplitude Perturbation Quotient sAPQ* /%/ and *Peak-to-Peak Amplitude Coefficient of Variation vAm* /%/ [5].

Voice break related measurements: *Degree of Voice Breaks DVB* /%/ [15] - the ratio of the total length of areas representing voice breaks to the time of the complete voiced sample; and *Number of Voice Breaks NVB.* The criteria for voice break area can be a missing impulse for the current period or an extreme irregularity of the pitch period.

Sub-harmonic components related measurements: *Degree of sub-harmonics DSH* /%/- the ratio of the number of autocorrelation windows with incorrect sub-harmonic period classification to the total number of autocorrelation windows; and *Number of Sub-Harmonic Segments NSH*

Voice irregularity related measurements: *Degree of Irregular Vocalization DUV* /%/ [15]- the ratio of the number of autocorrelation windows classified as unvoiced to the total number of autocorrelation windows; and *Number of Unvoiced Segments NUV.*

4.NOISE EVALUATION

The analysis of the voice signal in the frequency domain provides another approach to the evaluation of its irregularity (noise). The amount of in-harmonic spectral components correlates to the perception of hoarseness of the pathological voice [16]. *To* evaluate the level of noise components and separate the turbulent noise correlating to the intensity of the function *w(n)*, a pitch-synchronous frequency-domain method was developed. The following parameters are extracted: *Noise to Harmonic Ratio NHR*- a general evaluation of the noise presence in the analyzed signal (including amplitude and frequency variations, turbulence noise, sub-harmonic components and/or voice breaks); *Voice Turbulence Index VTI*- mostly correlating with the turbulence components caused by incomplete or loose adduction of the vocal folds; and *Soft Phonation Index SPI*- an evaluation of the poorness of high-frequency harmonic components that may be an indication of loosely adducted vocal folds during phonation.

The algorithm consists of the following general procedures:

1. Election of two groups of windows of 81.92 ms (4096 points) of the voice signal. The first group includes a sequence of windows of the voiced areas in the analyzed signal with a half window overlap. The second group includes four non-contiguous windows, where the frequency and amplitude perturbations are the lowest for the signal.

2. For every window in both groups the following steps apply: low-pass filtering (cutoff 6000Hz, order 22, Hamming window), downsampling to 12.5kHz and conversion of the real signal into analytical one using Hilbert filtering; computation of the power spectrum of the window using a 1024-points Complex Fast Fourier Transform (FFT) on the analytical signal; calculation of the average fundamental frequency within the current window from the time-domain analysis data and synchronous harmonic/in-harmonic separation; computation of the current window's *NHR*, *SPI* and *VTI*. *NHR* is a ratio of the in-harmonic energy in the range 1500-4500Hz to the

harmonic spectral energy (70-4500 Hz) and *SPI* is a ratio of the lower-frequency (70-1600Hz) to the higher-frequency (1600-4500Hz) harmonic energy for the first group of windows. *VTI* is a ratio of the spectral in-harmonic high-frequency energy (2800-5800Hz) to the spectral harmonic energy (70-4500Hz) for the second group of windows.

3. Computation of the average values of NHR, SPI and VTI.

5.TREMOR ANALYSIS

The pitch extraction process yields the amplitude and frequency demodulation curves of the voice signal. These curves contain information about the long-term amplitude and frequency variability (tremor) of the voice signal [17]. Methods for frequency and amplitude tremor analysis are developed. The algorithm for frequency tremor analysis includes the following steps:

1. Division of the *Fo*-data resulting from pitch extraction into windows of 2 sec. length with 1 sec. step overlap.

2. Application of the following procedures to every window: low-pass filtering of the *Fo*- data (cutoff 30Hz) and downsampling to 400Hz; calculation of the total energy of the resulting signals; subtraction of the DC-component and computation of the autocorrelation function on the residual signal; division of the autocorrelation data by the total energy and accumulation of the results from every window. The maxima of the resulting autocorrelation curve show the intensity and frequency of the long-term (up to 30Hz) frequency-modulating components.

3. Calculation of the *Fo-Tremor Intensity Index FTRI /%/*- the value of the global maximum of the average autocorrelation curve and the corresponding position *Fo-Tremor Frequency Fftr /Hz/*

The same method applies for computation of the *Amplitude Tremor Intensity Index ATRI /%/* and the *Amplitude-Tremor Frequency Fatr /Hz/* from the peak-to-peak amplitude data resulting from pitch extraction.

6.APPLICATION

Based on the model and the methods described above a *Multi-Dimensional Voice Program MDVP* was developed utilizing the *Computerized Speech Lab* (CSL) model 4300 (Kay Elemetrics Corp.). CSL, a hardware/software system which uses an MS-DOS based computer as host, includes signal conditioning, 16-bit A/D converters, dual digital signal processors (DSP16A & TMS32025) and support peripherals. The MDVP system computes a set of 33 acoustic voice parameters in about 16 seconds and provides flexible routines for graphical representation of the results[Fig.2-3]. Also a user-upgradable voice database allows automatic comparison of the current results with different nosological groups.

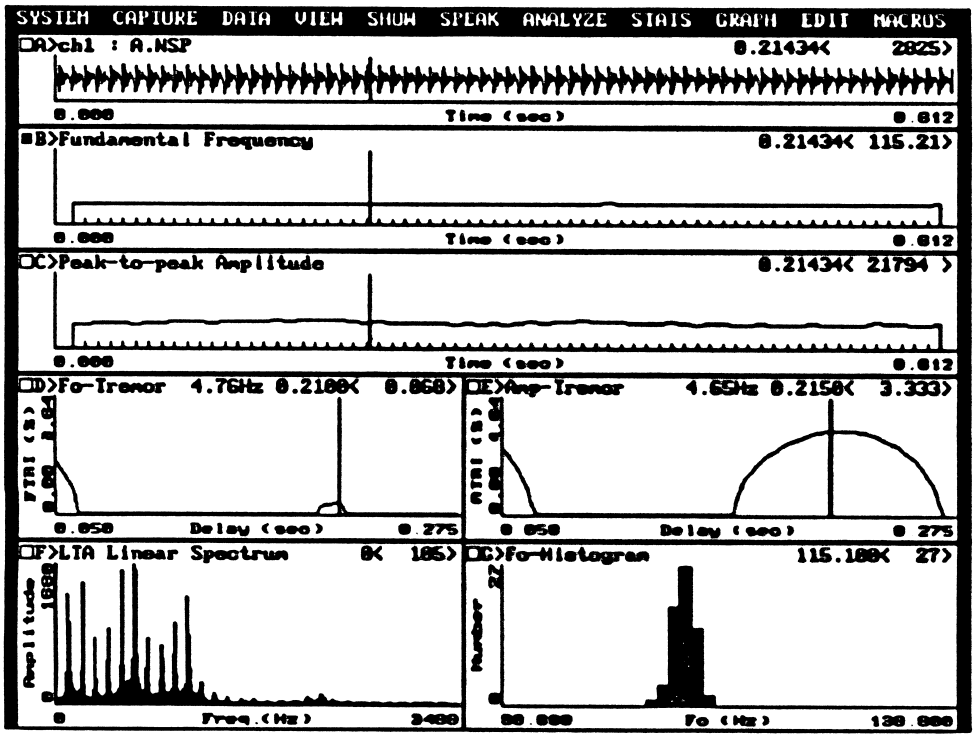


Fig.2: MDVP-Display of the voice waveform (view A), period-to-period fundamental frequency (B), peak-to-peak amplitude (C), *Fo*-tremor (D) and amplitude tremor (E) autocorrelation curves, long-term average linear spectrum of the voiced areas of the signal (F) and histogram of the distribution of *Fo* (G).

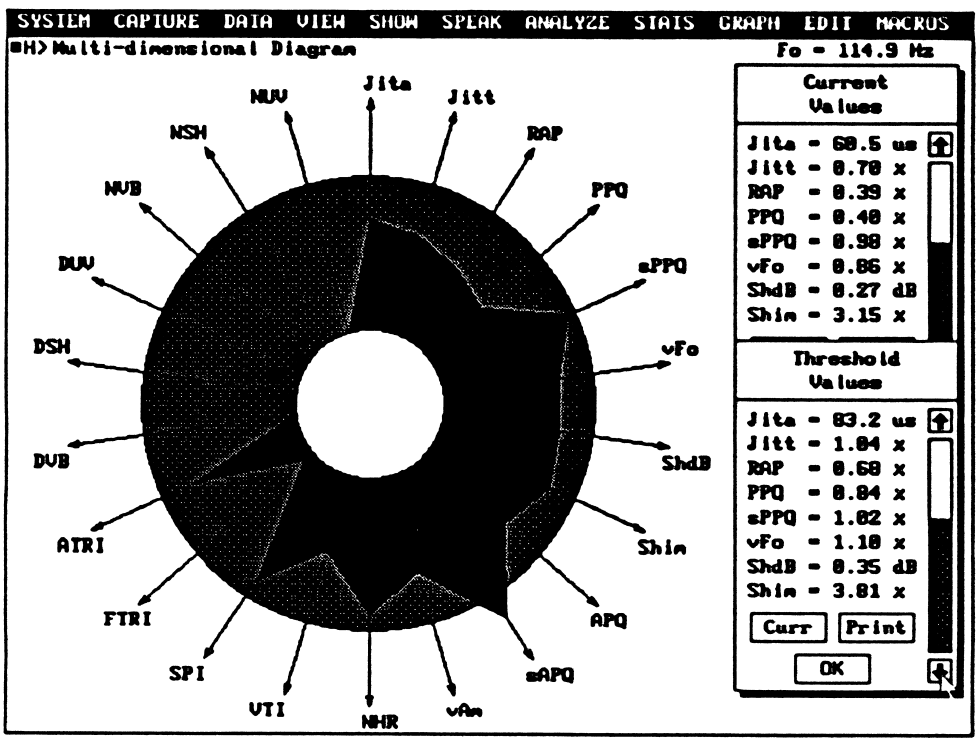


Fig.3: Multi-Dimensional Diagram display of the acoustic parameters. The area within the circle shows the normative threshold range and the polygon - the currently computed values.

In order to extract the normative threshold values of the acoustic parameters sustained phonation of the vowel 'a' of 15 persons (7m,8f) with normal voice production and of 53 patients (25m,28f) with laryngeal diseases were analyzed using the MDVP system. The following nosological groups were included in the study: laryngeal cancer, benign neoplasms, chronic laryngitis, functional dysphonia and paralysis of a recurrent nerve. The computed normative threshold values for this database are:

Frequency perturbation measurements:

Jita	Jitt	RAP	PPQ	sPPQ(55p)	vFo
83.2 us	1.04 %	0.68 %	0.84 %	1.02 %	1.10%

Amplitude perturbation measurements:

ShdB	Shim	APQ	sAPQ(55p)	vAm
0.35 dB	3.81 %	3.07 %	4.23 %	8.20 %

Voice break, sub-harmonic and voice irregularity measurements:

DVB	DSH	DUV	NVB	NSH	NUV
0 %	0 %	0 %	0	0	0

Noise and tremor evaluation measurements:

NHR	VTI	SPI	FTRI	ATRI
0.19	0.061	14.12	0.95 %	4.37 %

The normative values may vary depending on the nosological groups included in the specific study. A separate database is recommended to be selected or created for different applications.

7.FILE FORMAT

The format of sampled data files used by Kay Elemetrics Corp. was developed to meet the requirement for a single file that would contain any information that may be associated with a piece of sampled data and could be expanded to include additional features as those were incorporated into the program without rendering previous data files obsolete. A single file is advantageous because it keeps all information about a recording in one file. Separate files to describe a recording can be confusing and inadvertently separated. This file format is very flexible and is designed to be changeable to accommodate future requirements. Under exploration, for example, is the inclusion of videostroboscopic images with the file so that acoustic and images of the vocal cords can be viewed in synchronization with spectrograms and waveform displays. This new capability, unforeseen when the CSL was first developed, can be accommodated with the CSL file format without rendering previous data files obsolete.

Additionally, it was necessary that the format could be readily identified by any program attempting to read the file to determine that the file was, in fact, an appropriate sampled data file.

Toward these ends, a format made up of a number of nested named data BLOCKS was developed. The specification may be expanded by defining additional BLOCK types to accommodate new features and identifiability is provided since the name and placement of each BLOCK is specified for the file format and may be quickly checked as the file is read. This means, among other things, that it is not necessary to specify a particular filename extension in order to identify the file type to a program so that the extension may be put to better use as a classification aid for the user if desired.

A sampled data file which conforms to this specification contains the string "FORM" as the first 4 characters in the file (the FILE TITLE), followed by a BLOCK containing all data for the file. The BLOCK following the FILE TITLE may (and most certainly will) contain one or more nested BLOCKS. An example of Kay Elemetrics data file structure is shown of Fig.4.

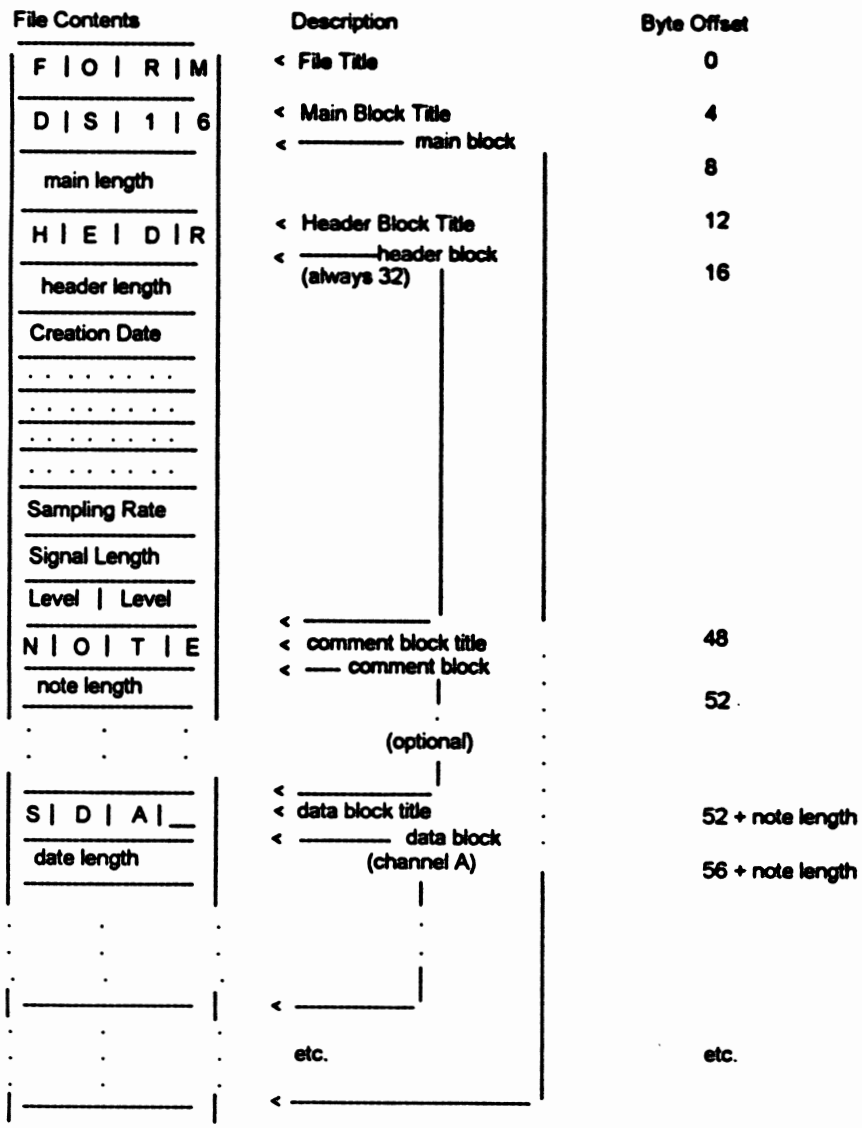


Fig.4: Kay Elemetrics Data File Structure Example.

Currently Kay's NSP data file format used in CSL can accommodate the following information: creation date, time and title, sampling rate, signal length, signal levels for each channel, sampled data from up to four channels, IPA phonetic transcription, named tags and voiced impulse markers for each channel, palatometric data and a comment field. Under concideraration is the inclusion of synchronous videostroboscopic images, signals associated with swallowing, patient's case history data, clinical evaluation and acoustic analysis results. Also the format is intended to accommodate several channels of data with different sampling rates.

The NSP format is very similar and easily convertible to RIF format, which is supported by Microsoft as a multi-media format. The products from the CSL-family support also input and output to several other file formats as TIMIT, ILS, DAT-tape, binary without header and flexible generic binary formats with header set by the user.

REFERENCES

[1]. Fant, C.G.M. *Acoustic theory of Speech Production*. The Hague: Mouton 1960.

[2]. Davis, S. *Acoustic Characteristics of Normal and Pathological Voices. Speech and Language Research and Theory.* Academic Press. N.J. 1979.

[3]. Titze, I., Baken, R., Herzel, H. Evidence of chaos in vocal fold vibration. *Vocal Fold Physiology.* Edited by Ingo Titze. Singular Publishing, USA. 1993.

[4]. Klatt, D.H., Klatt, L.C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J.Acoust.Soc.Am. 87, (2), 820-836, 1990.

[5]. Hirano, M. *Clinical Examination of Voice.* Springer Verlag. Vienna. 1981.

[6]. Kent, R. Vocal Tract Acoustics. Journal of Voice, Vol.7, No.2, 97-117, 1993.

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

[8]. Koike,Y, Takahashi,H.,Calcaterra,T. Acoustic measures for detecting laryngeal pathology. Acta Laryngol. 84, 105-117, 1977.

[9]. Deliyski, D. *Digital Processing of Voice Signals in the Diagnosis of Laryngeal Diseases.* Doctoral Dissertation, Bulgarian Academy of Sciences, Institute of Industrial Cybernetics and Robotics, Sofia, Bulgaria /in Bulgarian/ 1990.

[10]. Hess, W. *Pitch Determination of Speech Signals. Springer Verlag.* N.Y. 1983.

[11]. Rabiner, L. On the use of autocorrelation analysis for pitch detection. IEEE Trans. ASSP. Vol.ASSP-22, No.3, 1974.

[12]. Deliyski, D. Investigation of the autocorrelation function characteristics in pathologic voice signal analysis. 3-th International Conf. on Statistical Theory of Communications STS'88, Varna, Bulgaria, pp.17 /in Russian/, 1988.

[13]. Pinto, N., Titze, I. Unification of perturbation measures in speech signals. J.Acoust.Soc.Amer. 87, (3), 1278-1289, 1990.

[14]. Hillenbrand,J. Perception of aperiodicities in synthetically generated voices. J.Acoust.Soc.Amer. 83(6),2361-2371,1988.

[15]. Nikolov, Z., Deliyski D., Drumeva L., Boyanov B. Computer system for diagnostics of pathological voices. in Proc: XXI-st Congress International Association of Logopedics and Phoniatrics. Prague, Czechoslovakia, Vol.1, 973-976, 1989.

[16]. Kasuya, H., Ogawa Sh., K.Mashima K., Ehibara S. Normalized noise energy as an acoustic measure to evaluate pathologic voices. J.Acoust.Soc.Amer. 80, (5), 1986.

[17]. Winholtz W., Ramig L. Vocal tremor analysis with the vocal demodulator. J. Speech Hearing Res., Vol.35, 562-573, 1992.

Minimizing the Effect of Period Determination on the Computation of Amplitude Perturbation in Voice

Yingyong Qi, Bernd Weinberg, and Ning Bi
Department of Speech and Hearing Sciences
University of Arizona

Wolfgang J. Hess
Institut für Kommunikationsforchung und Phonetik
Universität of Bonn

May 4, 1994

Abstract

Current methods of computing amplitude perturbation of voice depend upon being able to accurately determine fundamental period. In this paper, we describe two methods of estimating amplitude perturbation of voice which do not depend on being able to accurately determine the boundaries of fundamental periods. In both of these methods, amplitude perturbation is computed as the variance of an ensemble of periods after these periods have been aligned in time. In one method, time alignment is accomplished using zero-phase transformation. In the second method, an unconstrained dynamic programming procedure is used. Accuracy of estimating amplitude perturbation by these two methods is evaluated with synthetic and natural voice signals and is also compared with estimation using zero-padding based time alignment. The unconstrained dynamic programming method is shown to provide accurate estimation of voice amplitude perturbation over a variety of signal conditions.

I. Introduction

Laryngeal diseases and disorders may cause disturbances in the voice signal. One significant disturbance is the presence of noise (Horii, 1980; Yumoto et al., 1982; Hillenbrand, 1987). The level of noise present in human voice often is difficult to quantify, in part, because the voice signal is complex and quasi-periodic (Kasuya et al., 1986; Muta et al., 1988; Qi, 1992). Because of the complex, quasi-periodic nature of human voice, many well-defined concepts in signal processing may not be directly applicable to the analysis of human voice signals. For example, the fundamental frequency of a periodic signal, $f(t) = f(t + nT)$, $n \in \mathbf{I}$, is defined as $\frac{1}{T}$. In theory, this definition cannot be applied to human voice signals because these signals are not truly periodic. Similar problems exist for the amplitude of voice signals as well. For example, it is known that the amplitude of a sinusoid refers to the maximum positive or negative excursion of the sinusoid from zero. The amplitude of a complex, periodic signal often refers to the amplitude of each sinusoidal component of the complex signal (Oppenheim and Schafer, 1989). The amplitude of a complex, quasi-periodic voice signal is not well identified. In this paper, we used the term fundamental period to refer to the duration between acoustic events that correspond to one cycle of vocal fold or voice source vibration. Fundamental frequency (f0) is the inverse of the fundamental period. The term amplitude is used to refer to the value of the voice signal at any instant in time. Amplitude perturbation refers to the total random variation in amplitude within one fundamental period.

The level of amplitude perturbation can be computed relatively easily as the ensemble variance of several periods, when all periods have the same length (Papoulis, 1984). The periods of human voices do not have the same length. Time-normalization of periods is necessary to compute the ensemble variance of voice signals. One method of time normalization is zero-padding in which zeroes are added to every short period.

Zero-padding can be used when the level of f0 perturbation is relatively small (Yumoto et al., 1982). When the perturbation in fundamental frequency is relatively large, for example, in pathological voices, the zero-padding normalization method should be used because the computed variance in amplitude will be significantly inflated by f0 perturbation. By way of example, two periods of a voice signal are shown in Figure 1a and their difference is shown in Figure 1b. As can be seen, the amplitude differences between these two periods are primarily due to the difference in temporal structure of the signals. If one period is compressed or stretched, the amplitude perturbation or difference between the two periods is negligible (see Figures 1c and 1d).

One of us has recently suggested that voice amplitude perturbation should be estimated as the ensemble variance in amplitude after all periods are optimally aligned in time (Qi, 1992). In this earlier work, optimal time-alignment of fundamental periods was accomplished using an end-point-constrained, dynamic programming procedure, in which the end-points of each period were aligned first, i.e., prior to optimal time alignment of every point within a period. This method of estimating amplitude perturbation was shown to be highly accurate even when relatively large f0 perturbations were added to voice signals. An assumption underlying this method of time-normalization is that the boundaries of each fundamental period can be determined accurately.

More recently, we have been conducting research to define acoustic properties of voices characterized by the presence of larger than normal levels of perturbations. To accomplish this work, we sought to develop methods of estimating amplitude perturbation which do not depend upon being able to accurately determine the boundaries of fundamental periods. Two such methods are described and evaluated in this paper. In both of these methods amplitude perturbation is computed as the variance of an ensemble of periods following time-normalization of these periods. In one method, time-normalization is accomplished using zero-phase transformation. In the second

method, an unconstrained dynamic programming procedure is used to time-normalize signals (Rabiner et al., 1978).

II. Methods of Time Normalization

Time-normalization is used in the computation of amplitude perturbation to minimize error due to temporal mis-alignment of individual periods. Two major sources of temporal mis-alignment are the time-aliasing effect among periods and errors in period boundary determination (PBD). Time-aliasing refers to the influence, due to the *infinite* impulse response of the vocal-tract, of previous periods on the period under analysis (Oppenheim and Schafer, 1989; Verhelst, 1991). When all periods have the same length, the influence of previous periods is constant and would not alter the temporal structure of each period. When f0 perturbation exists, the influence of previous periods varies on a period-by-period basis, resulting in the imposition of variations in the temporal structure on each period. Errors in PBD also produce alterations in the temporal structure of each period.

A. Zero-Phase Transformation

One approach to minimizing the effects of time mis-alignment on the estimation of amplitude perturbation is to remove all phase-related information for each fundamental period. This can be accomplished using zero-phase transformation. A four step computational approach is used to accomplish zero-phase transformation:

- Identify approximately period boundaries of a voice segment.
- Compute period-synchronized, zero-padded Fast Fourier Transformation (FFT) for each period.

- Compute the magnitude spectrum and set the phase of each frequency component to zero.

- Inversely transform the zero-phased magnitude spectrum.

Because phase-related information is removed in zero-phase transformation, all frequency components of each fundamental period are aligned in time prior to the computation of amplitude perturbation. Sample synthetic signals before and after zero-phase transformation are shown in Figure 2 to illustrate that time-misalignment between signals can be removed by this process.

Upon first consideration, it might appear that zero-phase transformation is essentially a frequency-domain approach to the estimation of amplitude perturbation. This initial view suggests that the inverse transformation may not be necessary, i.e., the magnitude spectrum could be used directly to estimate amplitude perturbation. However, the durations of individual periods are not equal in human voice and the harmonic frequencies of the discrete magnitude spectrum would be expected to vary from period-to-period. This variation makes it difficult to use the magnitude spectrum directly for the estimation of amplitude perturbation. The inverse Fourier transform brings each period back in the time domain with the same length and the resulting computation of ensemble variance of the inversely transformed signals is simple and straightforward.

A potential drawback of the zero-phase transformation method is the numerical implementation of the Fourier transform. The FFT algorithm always assumes that the signal segment under analysis is periodic (Oppenheim and Schafer, 1989). With this assumption, random errors in PBD will not simply be shifts in the time origin. Rather, errors in PBD will produce changes in the cyclic pattern of the signal and degrade a period-synchronized FFT into a period-asynchronized FFT. It is well recognized that period-synchronized FFT provides more accurate spectral estimation of voice signals

than does period-asynchronized FFT (Kay, 1987).

B. Unconstrained Dynamic Programming

A second approach to minimizing the effect of time mis-alignment of periods on the estimation of amplitude perturbation is dynamic programming (DP) procedures. Dynamic programming optimally minimizes the differences between signals that are due to temporal mis-alignment (Nemhauser, 1966). Optimal implies that there will not be another temporal alignment that can produce a smaller difference between signals under a given set of conditions. The conditions of DP are often stated heuristically to facilitate the optimization process. For example, in a constrained DP approach, the end-points of signals cannot be shifted in time. In an unconstrained DP approach any points can be shifted in time to achieve optimal match between signals (Parsons, 1987). We felt the unconstrained DP approach should offer advantages for evaluating amplitude perturbation in voices, particularly when period boundaries are difficult to determine.

The algorithm for unconstrained DP (Brown and Rabiner, 1982) is similar to that for constrained DP (Qi, 1992), except for the processing of starting and ending points of each period. The process of time-normalization can be viewed as the search for an optimal matching path through the lattice of points (see Figure 3). The specific algorithm used in this work is briefly summarized below:

1. At the *ith* step in the horizontal direction, the lower limit and the upper limit for searching in the vertical direction were given by $\max(1, \frac{i \cdot N}{M} - \delta)$ and $\min(N, \frac{i \cdot N}{M} + \delta)$, respectively. These searching boundaries defined a polygon, shown in Figure 3.

2. Within these searching limits, the path for connecting each point (i, j) to previous points in the lattice was determined by minimizing the total cost (rms

differences) for reaching the current position. Specifically,

(a) Starting with all points on the searching border ($i = 1 \forall j$ and $j = 1 \forall i$). Because there were no predecessors, the squared difference between samples on these points was computed as the starting cost.

(b) Looping through all (i, j). In each loop, the costs from the current point (i, j) to the predecessors $(i-1, j)$, $(i, j-1)$, and $(i-1, j-1)$ were computed. The connection between point (i, j) and one of its predecessors was made such that the cost for making the connection plus the cost for reaching the particular predecessor was minimized. This minimum cost was stored as the cost for reaching point (i, j).

(c) When $i = M, j = N$, the search was terminated and a complete path could be retrieved from the point with minimum total cost on the ending border of the searching limits ($i = M, N - \delta \leq j \leq N$ and $j = N, M - \delta \leq i \leq M$).

3. The final amplitude difference between any two periods was equal to the minimum total cost on the ending border of the searching limits.

An example search is illustrated in Figure 3.

III. Experimental Procedures

To evaluate the use of time-normalization in the estimation of amplitude perturbation, the signal-to-noise ratios (SNRs) of synthetic and natural voices was computed The SNRs of three time-normalization methods — zero-padding (ZP), zero-phase transformation (ZPT), and unconstrained dynamic-programming (UDP) — were computed. SNR was defined as the ratio between the signal energy of the most representative period within a voice segment and the residue ensemble variance of all

other periods under analysis following time-normalization. The most representative period was the mean of the period ensemble when ZP or ZPT was used. The most representative period for UDP was the period with the minimum total rms distance to all other periods. The ensemble mean is not available when UDP is used for time-normalization (Qi, 1992).

A. Synthetic Voice Evaluation

The vowel /a/ was synthesized with a formant synthesizer. The synthesizer was a 5-pole, autoregressive digital filter whose coefficients were determined by 5 given pairs of formant frequencies and bandwidths (Rabiner and Schafer, 1978). The unperturbed excitation source to the synthesizer was an equally-spaced impulse train. The amplitude of the impulse was set to 1000. Controlled perturbations were superimposed on the impulse train and the synthesis was made by convolving the perturbed excitation source with the impulse response of the autoregressive filter. The sampling frequency of the synthesizer was 16 kHz. Twenty periods were synthesized for each SNR computation.

Amplitude perturbation was introduced by adding a zero-mean, Gaussian random noise to the impulse train. The level of the noise was controlled by the standard deviation of the Gaussian distribution, given as the percentage of the impulse amplitude. Fundamental frequency perturbation was introduced by adding a zero-mean, uniformly distributed random number to each period of the impulse train. The level of f_0 perturbation was controlled by the standard deviation of the random number generator, given as the percentage of the average period. Error in period determination was introduced by adding another zero-mean, uniformly distributed random number to the known location of each impulse after the vowel had been synthesized. The level of the error was controlled by the maximum of the random number generator, given in number of samples.

The effect of amplitude perturbation, f0 perturbation, and PBD error on SNR was determined by systematically varying the perturbation of one parameter, while holding the other parameters constant. To determine the effect of amplitude perturbation, the average f_0 (at 120 Hz and 220 Hz, respectively) and the degree of f_0 perturbation (5%) were held constant. The standard deviation of the Gaussian random noise was increased from 1% to 25% of the impulse amplitude (1000) in incremental steps of 5%. To determine the effect of fundamental frequency perturbation, the standard deviation of noise was held constant (5%). The standard deviation of f0 was increased from 1% to 25% of the fundamental period in incremental steps of 5%. To determine the effect of period boundary determination, the standard deviation of f0 perturbation was set to 1%, 5%, and 10% of the fundamental period, and the standard deviation of the noise generator was set to 1%, 5%, and 10% of the impulse amplitude, respectively. The maximum of the random number generator for producing PBD error was varied from 0 sample to 10 sample in incremental steps of 1 sample.

B. Natural Voice Evaluation

Natural voices were used to further evaluate SNR estimation. The natural voices were used only to determined the effect of PBD errors on the computed SNRs. The levels of amplitude and f0 perturbations are not controllable in such samples.

Sixteen, non-smoking, healthy adults (8 men and 8 women) provided voice samples. Each subject produced a sustained /a/ at a constant, comfortable intensity level for a duration of more than 1 second. The microphone (ASTATIC, CTM-80) was placed about 10 cm in front of the subject's mouth. A pistonphone (GENRAD, Minical-1987) was used to record a calibration tone prior to each recording session, and all recordings were made in a quiet room. The recorded productions were digitized into a computer (SUN, Sparc10/30) at a sampling frequency of 16 kHz and a

quantization level of 16 bits. The signal was passed through an anti-aliasing filter with a cut-off frequency of about 7.5 kHz prior to the digitization. A waveform editor (Speech Acoustic Lab, Ocean) was used to select a stable 20-period, segment for each subject.

The period boundaries of the selected voice segments were determined from the residue signal of linear predictive (LP) inverse filtering. The order of the LP filter was 12. The autocorrelation method and the Hamming window were used in the LP analysis. The window length was 256 points and the window step size was 128 points. The location of period boundaries was identified using a time-delayed, peak-picking algorithm. Time-delay was introduced to ensure that each maximum located was global within a given time bracket and demarcated boundaries of the fundamental periods. These period marks were assumed to be the correct period boundaries.

Error in PBD was introduced by adding a zero-mean, uniformly distributed random number to the absolute time locations of detected period boundaries. The level of PBD error was controlled by the maximum of the random number generator. This maximum varied from 0 sample to 10 sample, in incremental steps of 1 sample. The altered locations were used as the period boundaries for SNR computation. The SNRs were computed in the same manner as described earlier for synthetic voices.

A two-step procedure was used in the statistical analyses of the computed SNRs. First, a polynomial regression (3rd order) of SNR as a function of PBD error was made for each subject. Second, an analysis of variance (ANOVA) was undertaken to assess the effects of gender and method of SNR estimation on the coefficients of the regression polynomial. In the ANOVA, the dependent variables was the coefficients of the regression polynomial. The independent variables were gender group, method of SNR estimation, and the interaction between gender group and method of SNR estimation.

IV. Results and Discussions

The computed and known SNRs of the synthetic signals are plotted as a function of noise level in Figure 4, as a function of f0 perturbation in Figure 5, and as a function of PBD error in Figures 6, 7 and 8. For synthetic voice signals, ZPT- and UDP-based SNRs accurately define known SNRs over a wide range of signal conditions/perturbations. ZP-based SNRs significantly underestimate known SNRs (see Figure 4). ZPT- and UDP-based calculations of SNR were not significantly influenced by level of f_0 perturbation, whereas ZP-based SNR was significantly influenced by level of f_0 perturbation (see Figure 5). UDP-based calculations of SNR were also not significantly influenced by PBD error. ZPT-based calculations of SNR decreased slightly as level of PBD error increased, when perturbations in f0 and amplitude were small (see Figure 6). ZP-based calculations of SNR were significantly influenced by level of PBD error (see Figures 6, 7 and 8).

The computed SNRs for each of the 16 natural voices are plotted as a function of PBD error in Figure 9. The means and standard deviations of the regression parameters of the functions plotted in Figure 9 are tabulated in Table I. The ANOVA results of based on these parameters are provided in Table II.

The ANOVA results indicate that gender did not exert a significant effect on the regression parameters and that the interaction between gender and SNR estimation method was not significant. Method of SNR estimation did exert a significant influence on the regression parameters. *Post hoc* testing revealed that significant method differences in intercept and linear slope existed between ZP and ZPT and between ZP and UDP-based estimations. Significant differences in the quadratic and cubic terms were found between ZP and UDP and between ZPT and UDP-based estimations. ZP-based SNR functions exhibited a significantly smaller intercept and a significantly larger negative slope than ZPT- and UDP-based functions (see Figure

9). ZP- and ZPT-based functions exhibited significantly larger quadratic and cubic terms than UDP-based functions.

Three major findings emerge from these analyses. As expected, ZP-based SNRs generally did not provide accurate estimation of amplitude perturbations of synthetic and natural samples. Second, ZPT-based SNRs provided accurate estimation of amplitude perturbations of synthetic samples; however, estimation of amplitude perturbations present in natural voices measured by this method was significantly influenced by the level of PBD error. Third, UDP-based SNRs provided an accurate estimation of amplitude perturbations of synthetic and natural samples that was not significantly influenced by f0 perturbation and PBD error.

From a comparative point of view, computation of ZPT-based SNR is much simpler than UDP-based computation. ZPT computation requires only a forward and inverse Fourier transform and can be implemented using a fast algorithm. By contrast, UDP-based calculation is computationally intensive. The UDP method is, however, robust in the presence of f0 perturbation and PBD error, presumably because this method optimally minimizes temporal influences in the calculation of amplitude perturbation.

This work was motivated, in part, by our need to develop a method of measuring amplitude perturbation that did not depend upon accurate determination of period boundaries. In this paper and in prior work (Qi, 1992), we have attempted to demonstrate how time-normalization can significantly influence estimation of amplitude perturbation. One of our goals was to develop procedures that optimally separate temporal from amplitude variations in voice signals. UDP-based SNRs for normal voices and a wide range of synthetic voices have been shown to provide accurate estimates of amplitude perturbation that are independent of f0 perturbation and PBD error. Based on this observation, we are currently using the UDP-based method to estimate amplitude perturbation in both pathological and normal voices.

V. Acknowledgment

This work was supported, in part, by a grant from the National Institute of Deafness and Other Communication Disorders, DC01440, Analysis and Improvement of Alaryngeal Speech. We gratefully acknowledge the contributions of Patricia Jones Ph.D. to the statistical analyses of this work.

References

Brown, M. and Rabiner, L. (1982). An adaptive, ordered, graph search technique for dynamic time warping for isolated word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-30:535-544.

Hillenbrand, J. (1987). A methodological study of perturbation and additive noise in synthetically generated voice signals. *Journal of Speech and Hearing Research*, 34:448–461.

Horii, Y. (1980). Vocal shimmer in sustained phonation. *Journal of Speech and Hearing Research*, 23:202-209.

Kasuya, H., Ogava, S., Mashima, K., and Ebihara, S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. *J. Acoust. Soc. Amer.*, 80:1329.

Kay, S. M. (1987). *Modern Spectral Estimation.* Prentice-Hall, Englewood Cliffs, New Jersey.

Muta, H., Baer, T., Kikuju, W., Tervo, M., and Fukuda, H. (1988). A pitch-synchronous analysis of hoarseness in running speech. *J. Acoust. Soc. Amer.*, 84:1292–1301.

Nemhauser, G. (1966). *Introduction to Dynamic Programming.* Wiley, New York.

Oppenheim, A. and Schafer, R. (1989). *Discrete -Time Signal Processing.* Prentice-Hall, Englewood Cliffs, New Jersey.

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, 2 edition.

Parsons, T. (1987). *Voice and Speech Processing.* McGraw-Hill, New York.

Qi, Y. (1992). Time normalization in voice analysis. *J. Acoust. Soc. Amer.*, 92:2569–2576.

Rabiner, L., Rosenberg, A., and Levison, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-26:575–582.

Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals.* Prentice-Hall, New Jersey.

Verhelst, W. (1991). On the quality of speech produced by impulse driven linear systems. In *Proc. IEEE ICASSP*, pages 501–504.

Yumoto, E., Gould, W., and Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *J. Acoust. Soc. Amer.*, 71:1544–1550.

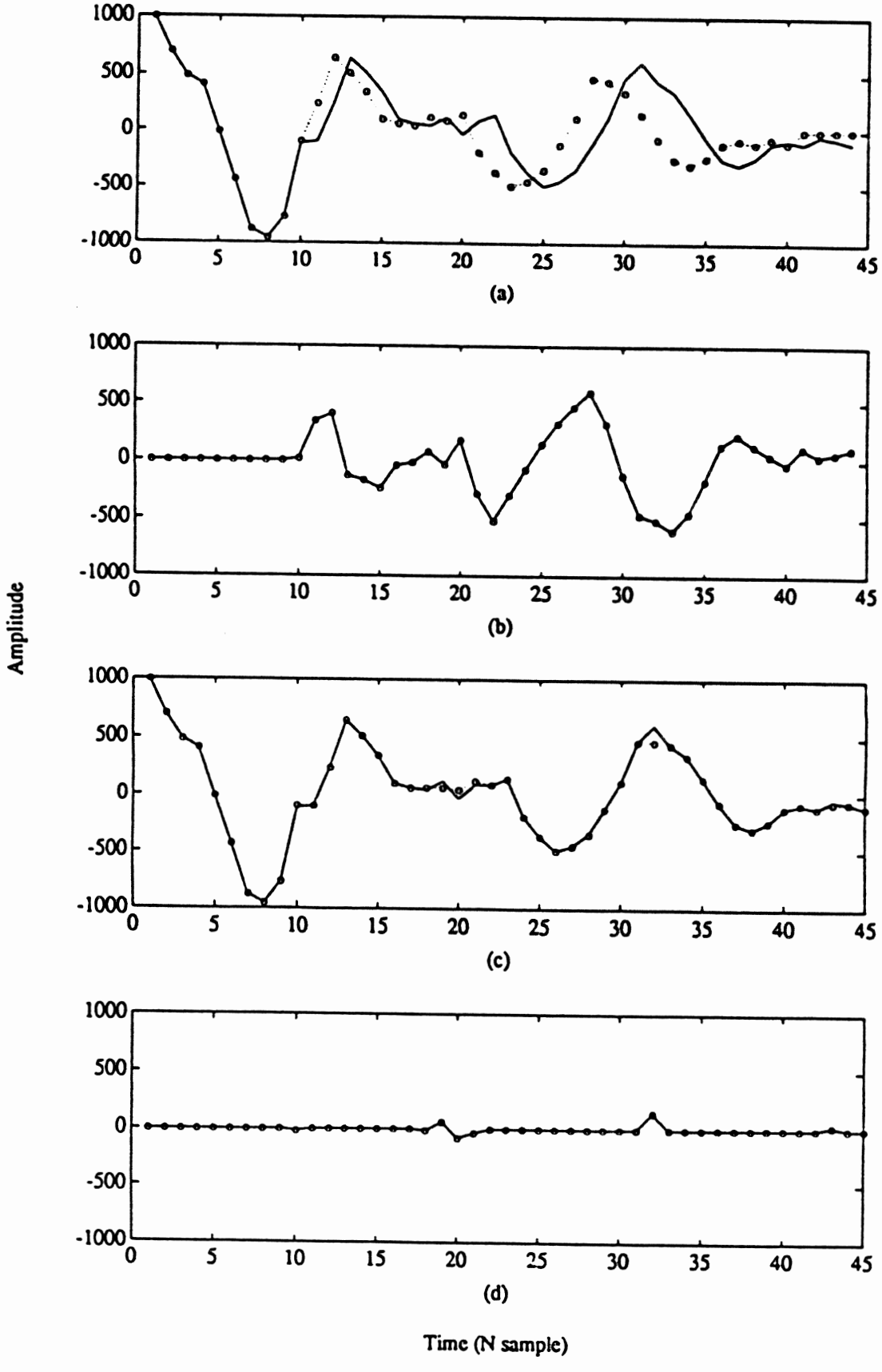


Figure 1. (a) A pair of waveforms. (b) Amplitude difference between waveforms. (c) The waveforms after optimal alignment in time. (d) The remaining difference between the waveforms.

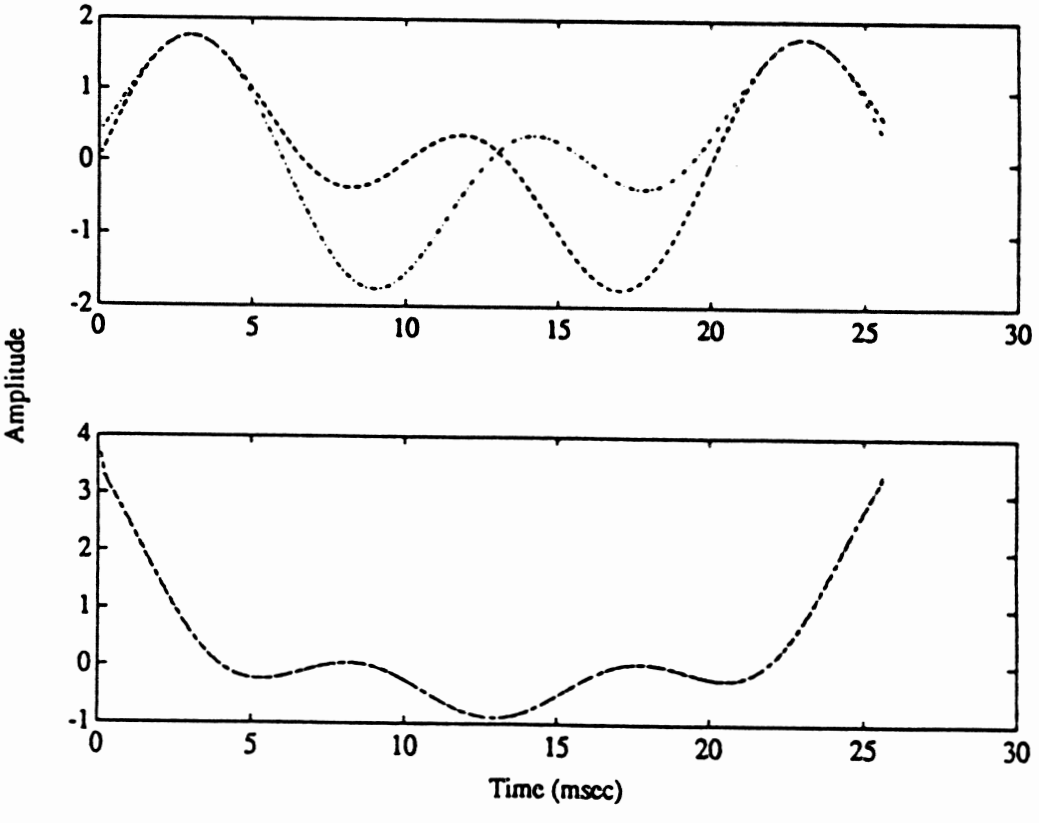


Figure 2. Sample signals before (top) and after (bottom) zero-phase transformation.

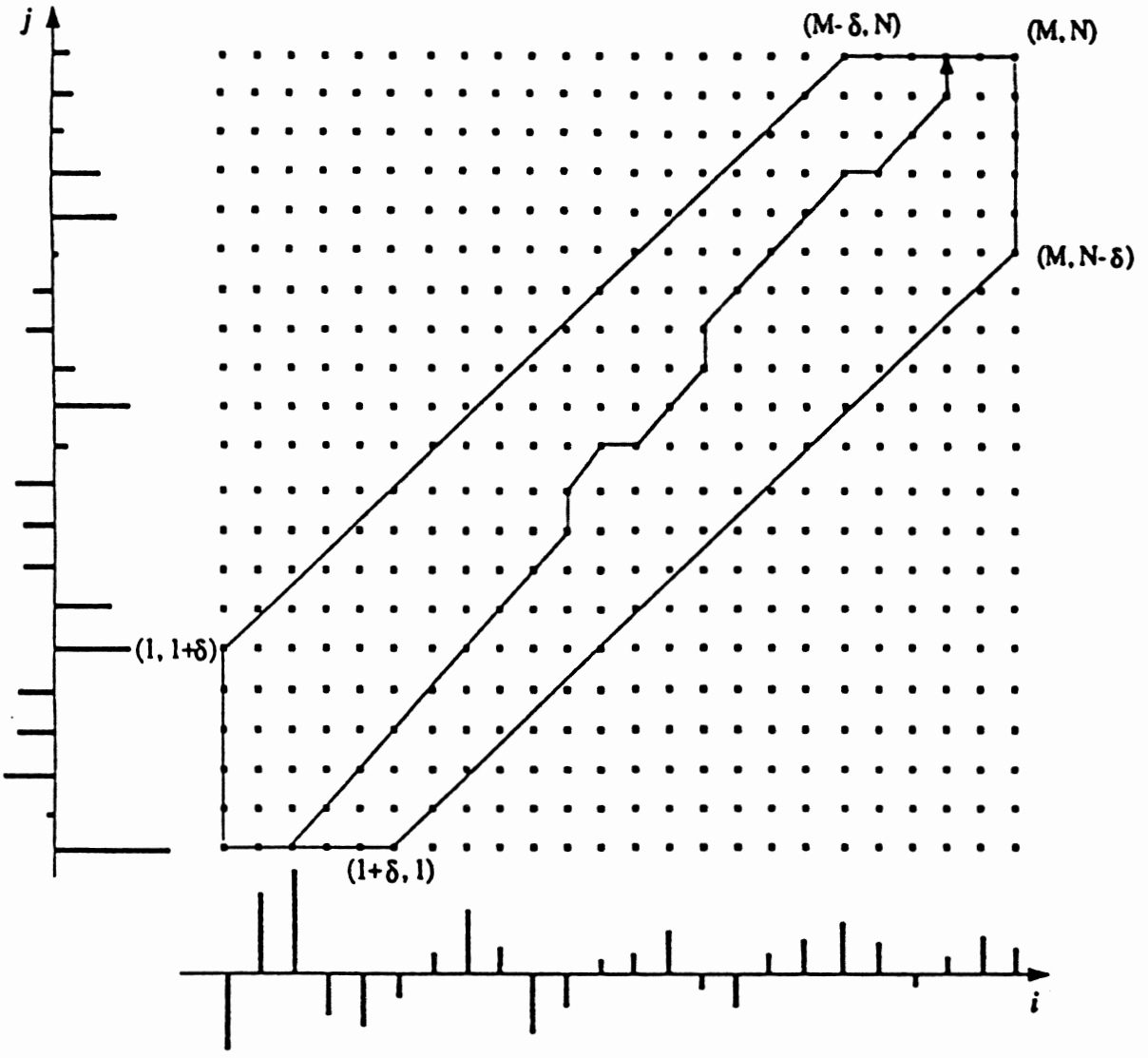
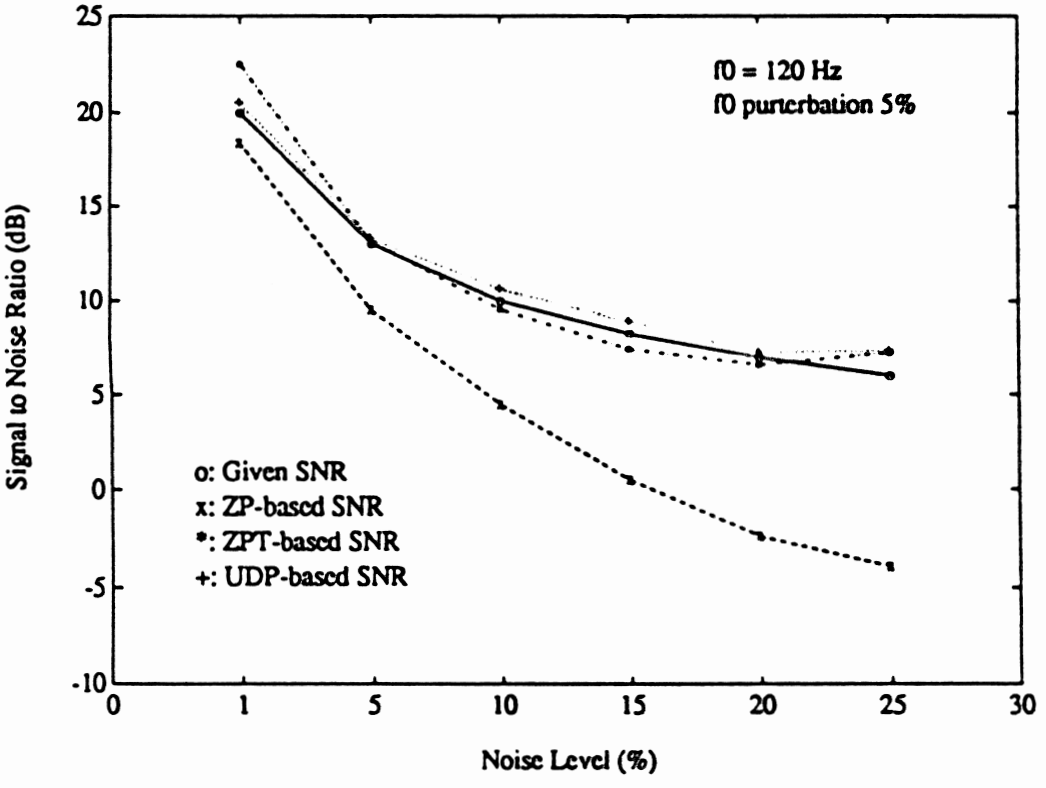


Figure 3. Illustration of the Unconstrained Dynamic Programming computation.



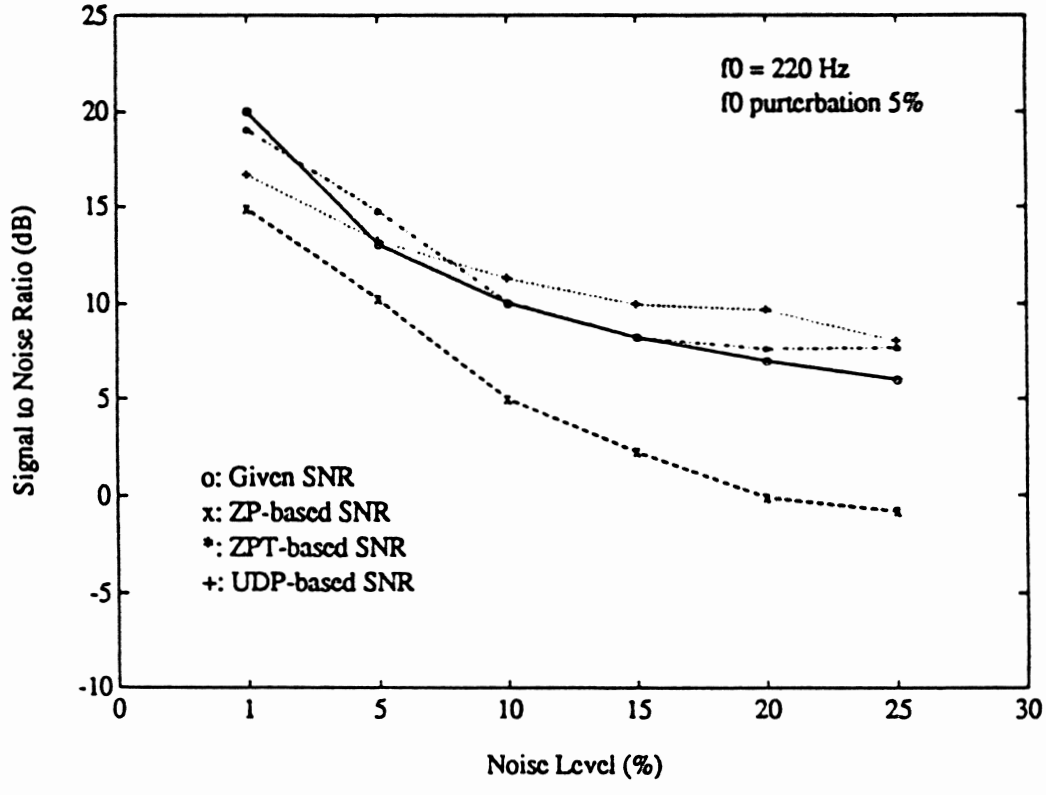
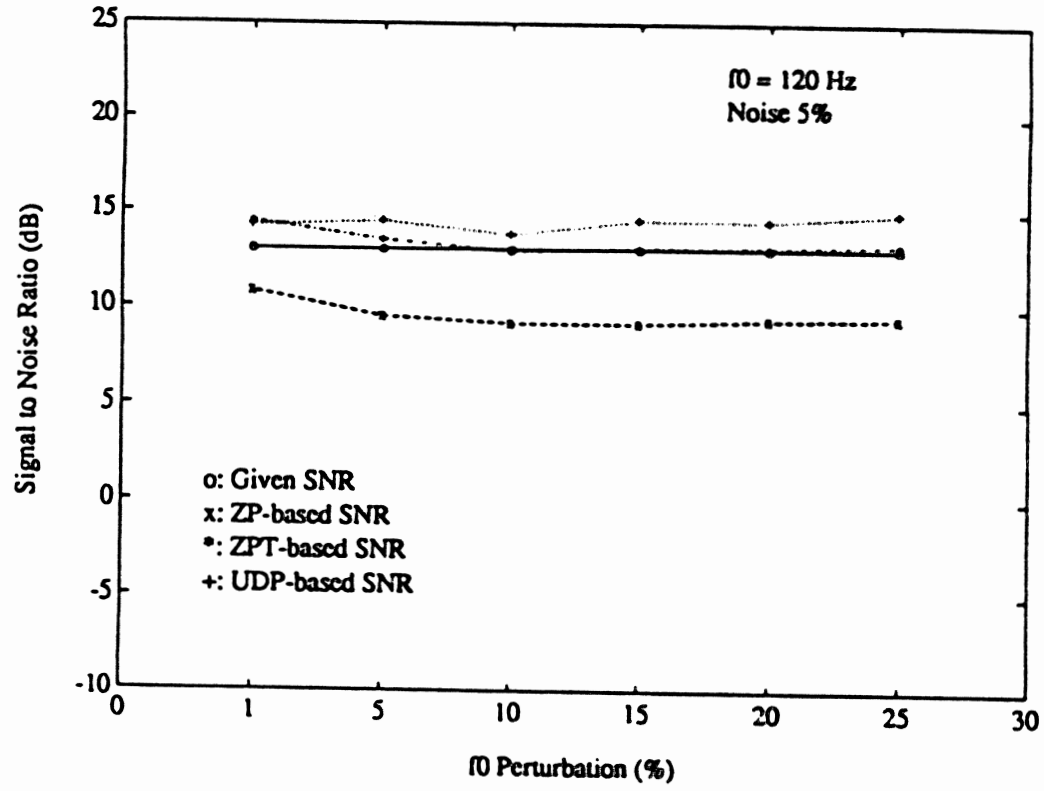


Figure 4. Signal-to-noise ratios as a function of relative noise level.



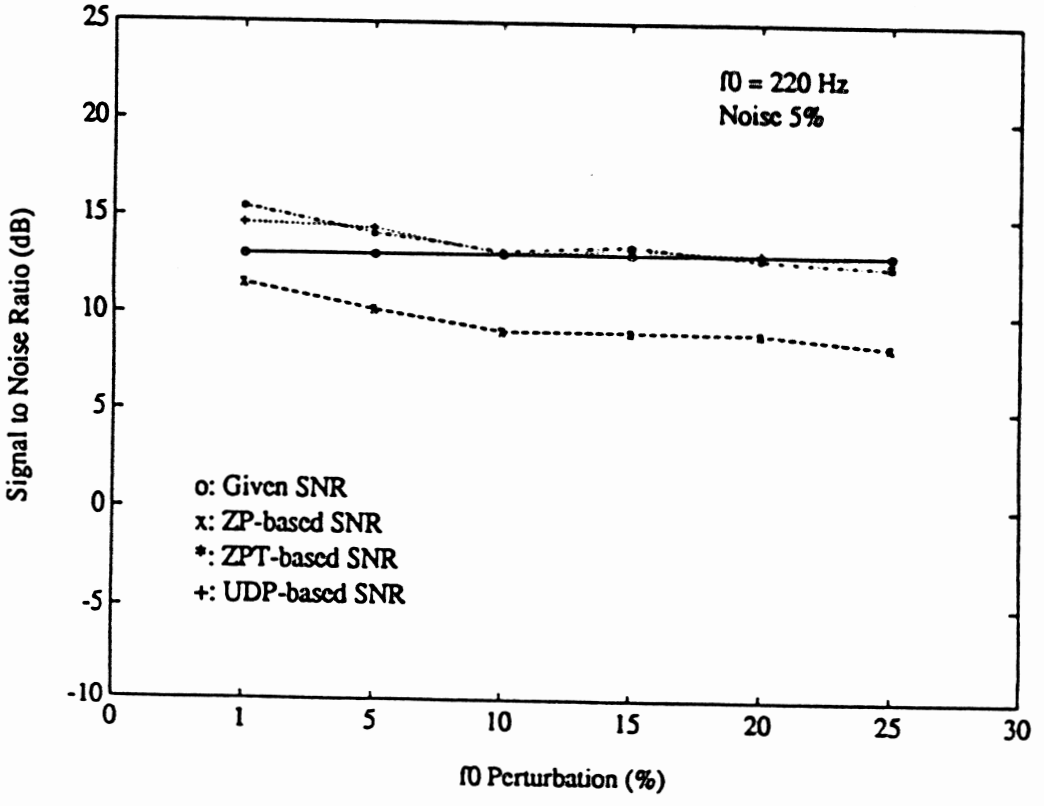
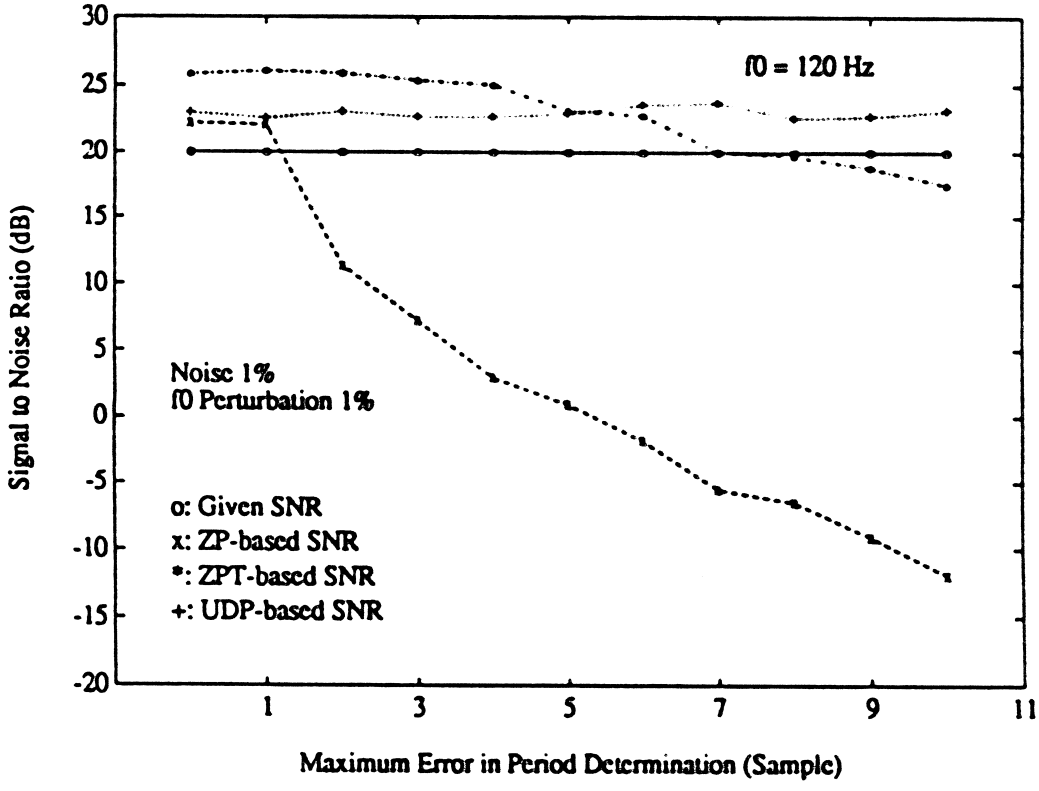


Figure 5. Signal-to-noise ratios as a function of fundamental frequency perturbation.



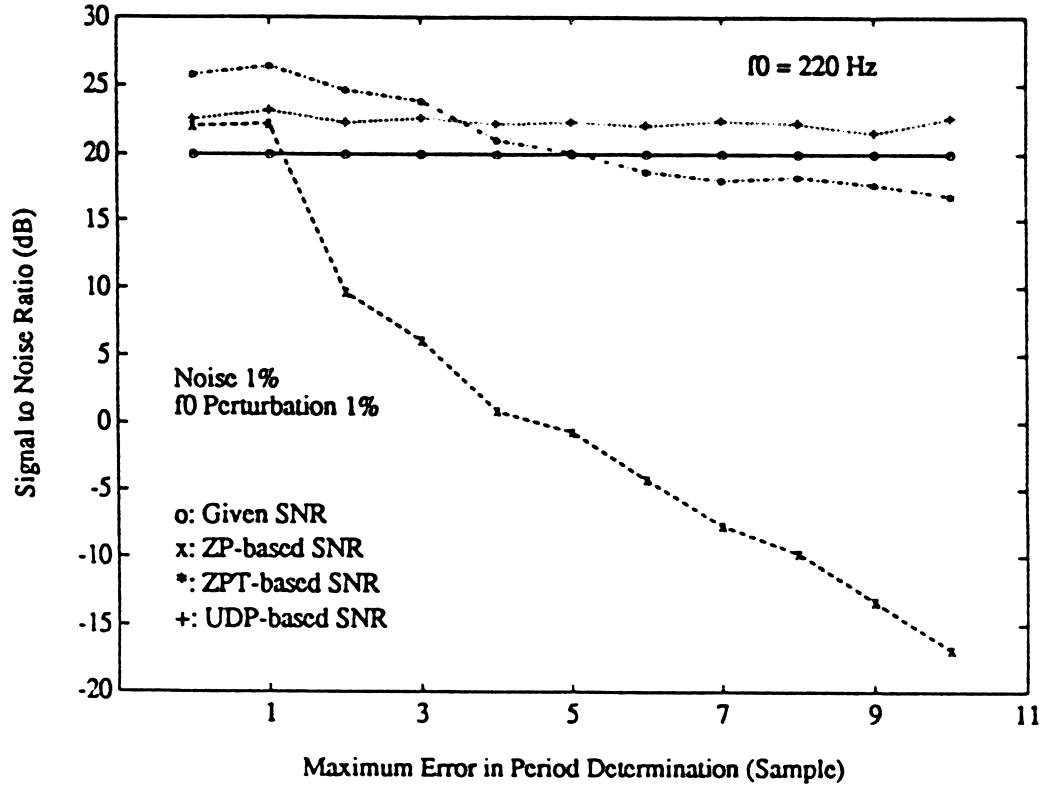
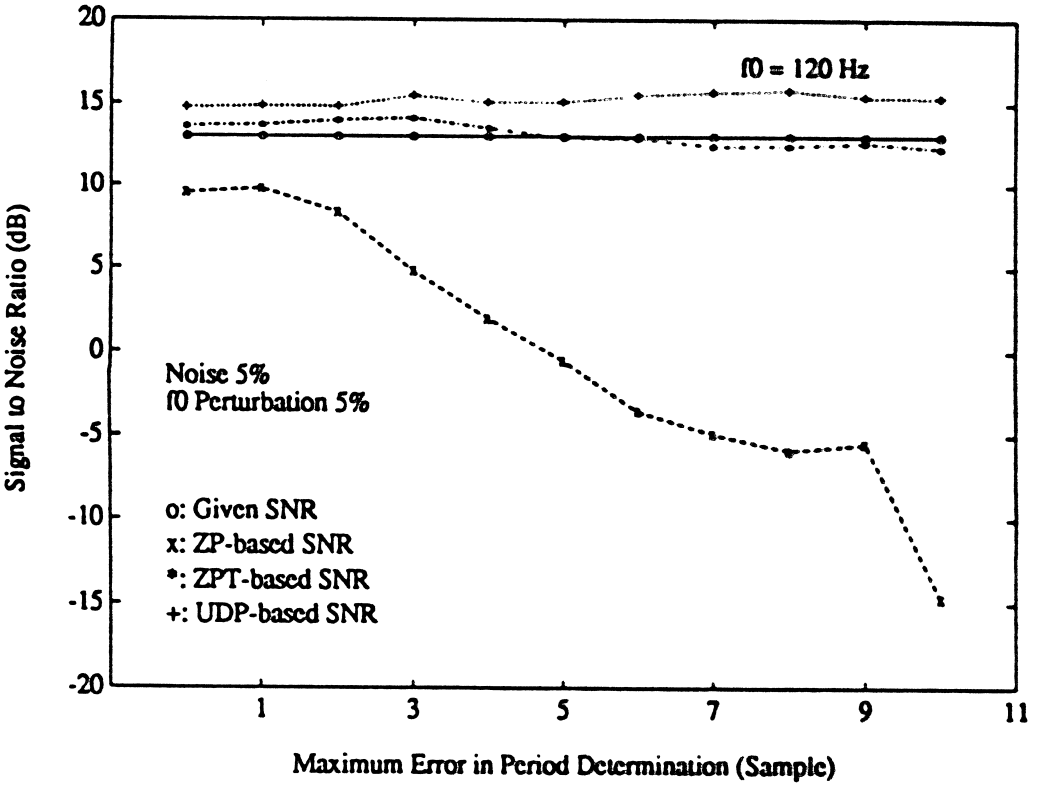


Figure 6. Signal-to-noise ratios as a function of period determination error when f0 and amplitude perturbation are at 1% level.



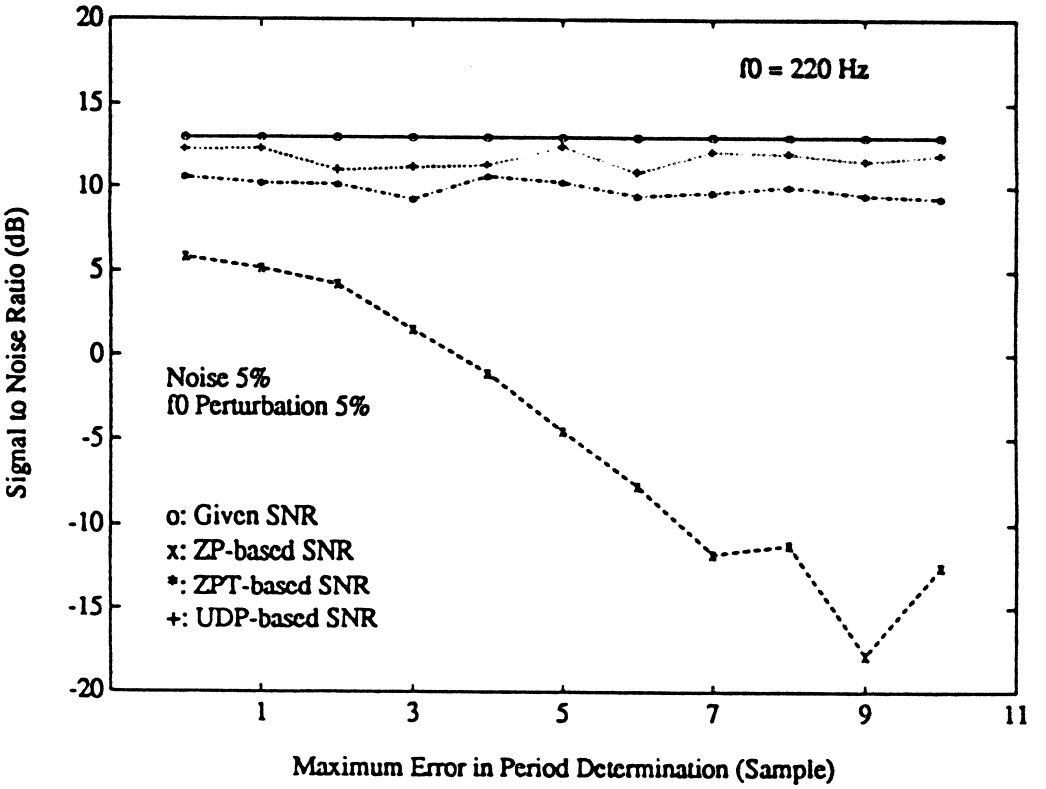


Figure 7. Signal-to-noise ratios as a function of period determination error when f0 and amplitude perturbation are at 5% level.

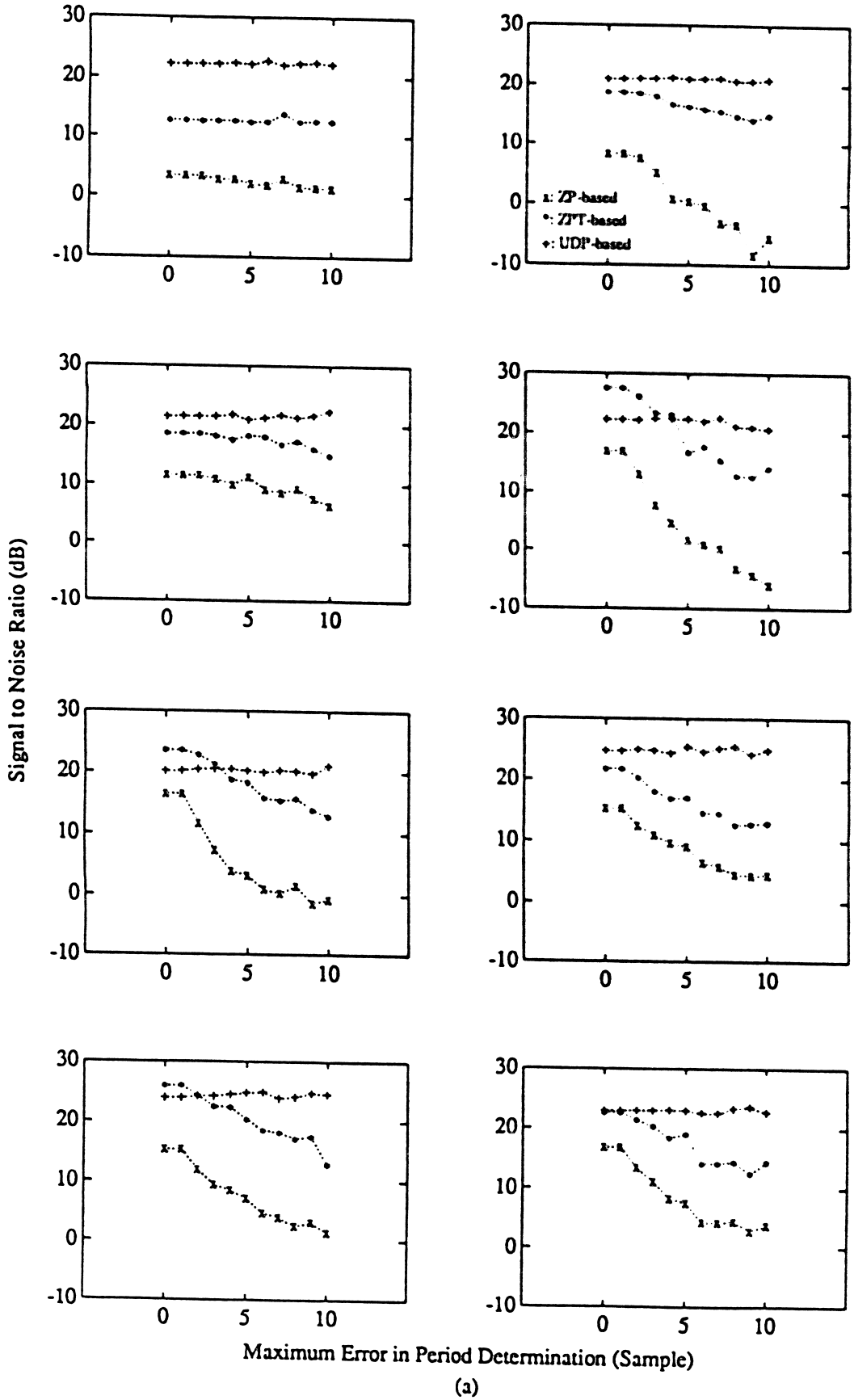
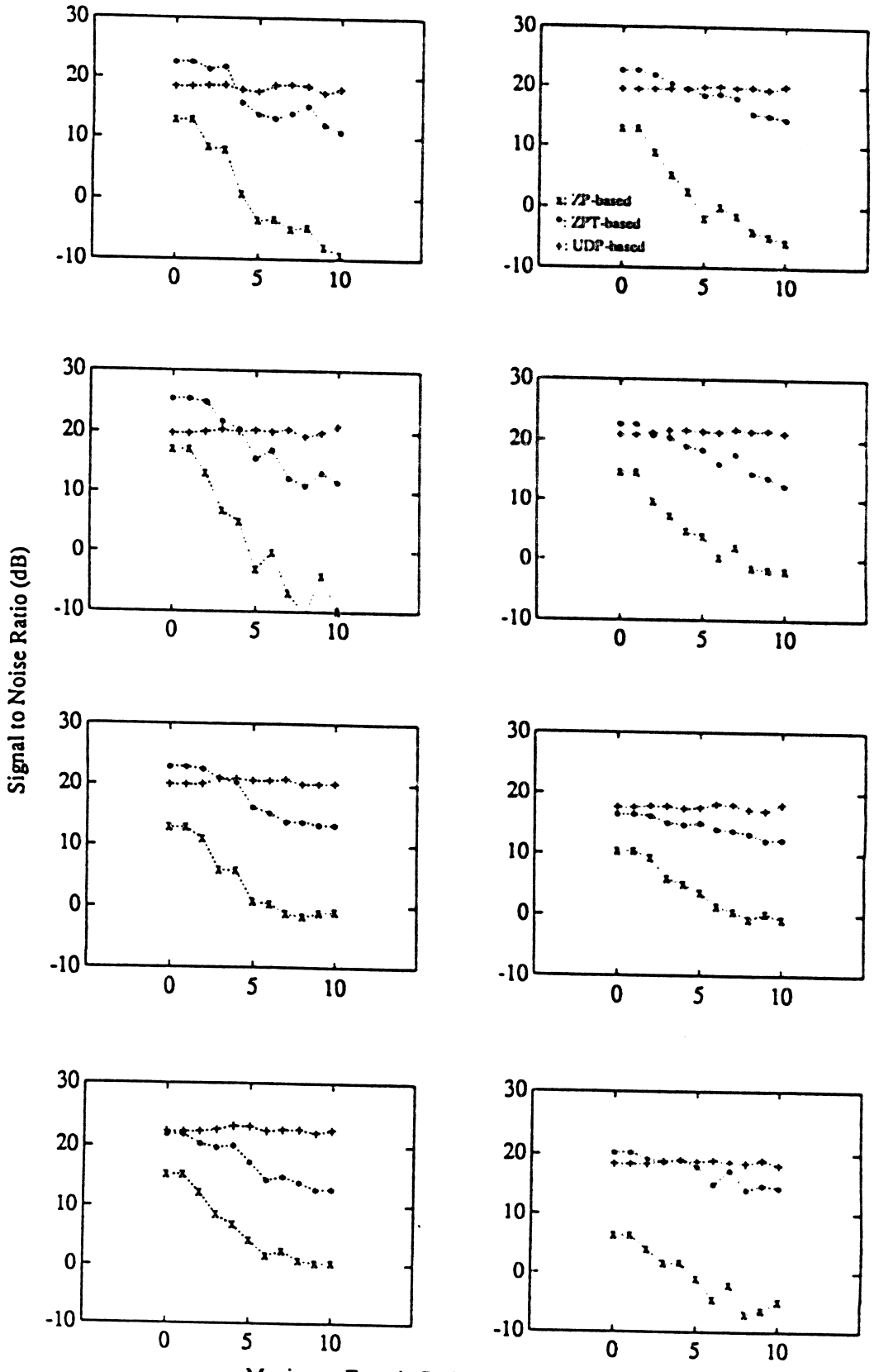


Figure 9. Signal-to-noise ratios as a function of period determination error for (a) male and (b) female subjects. Each small figure is for one individual subject.


x: ZP-based
•: ZPT-based
+: UDP-based
Signal to Noise Ratio (dB)
Maximum Error in Period Determination (Sample)

(b)

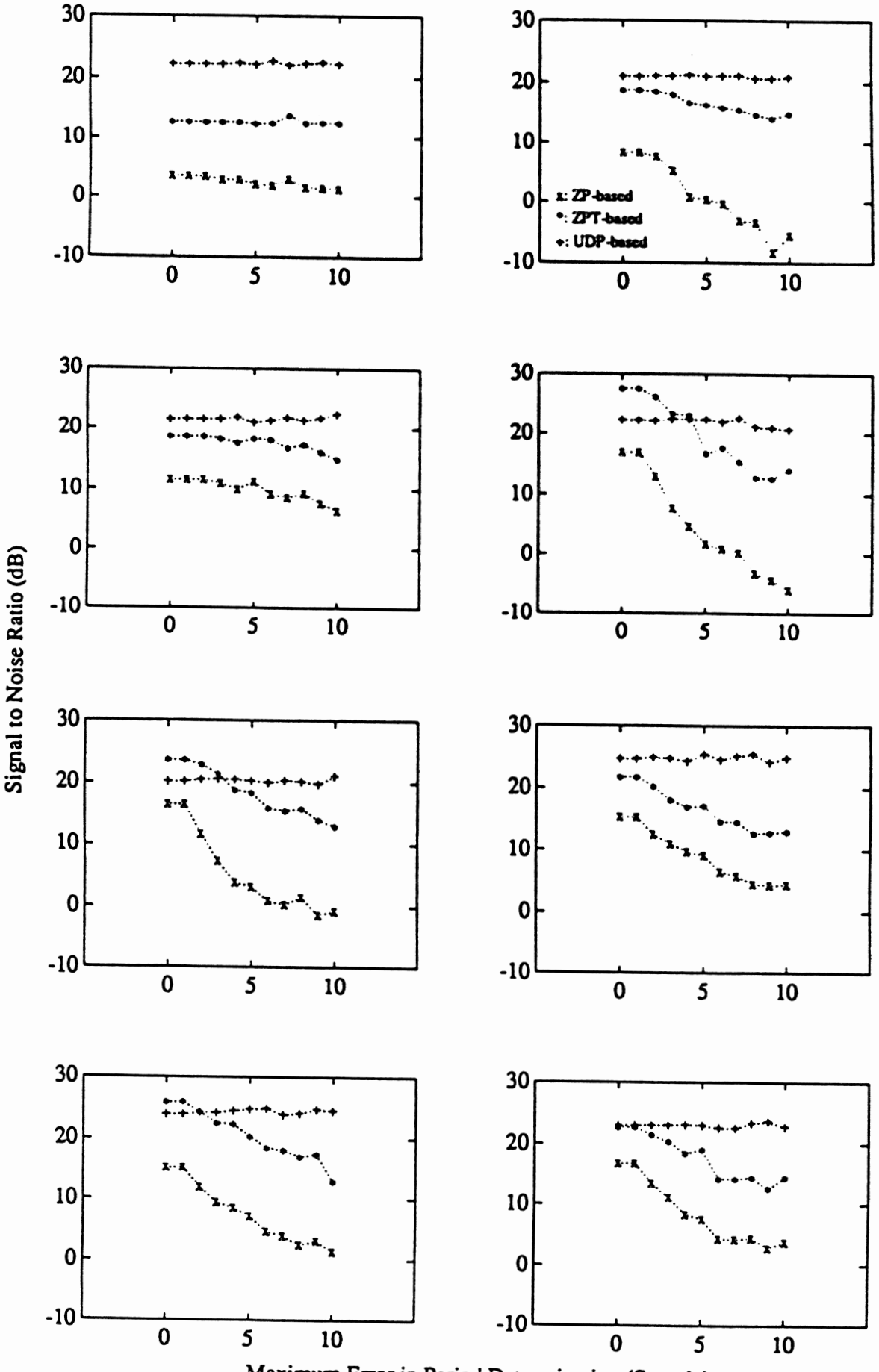
TABLE I. Results of Linear Regression Analysis

Regression Parameter	Gender Group	Normalization Method	Mean	SD
Intercept	Male	ZP	13.6889	5.2853
		ZPT	21.6647	4.8081
		UDP	22.2713	1.5360
	Female	ZP	13.6712	3.6189
		ZPT	22.1046	2.7006
		UDP	19.5100	1.4336
Linear	Male	ZP	-1.6421	1.4532
		ZPT	-0.3342	0.5430
		UDP	0.1561	0.2022
	Female	ZP	-1.9601	1.0239
		ZPT	-0.2434	0.6312
		UDP	0.2390	0.2380
Quadratic	Male	ZP	-0.0567	0.1667
		ZPT	-0.1707	0.2471
		UDP	-0.0343	0.0568
	Female	ZP	-0.2065	0.1973
		ZPT	-0.2291	0.2343
		UDP	-0.0373	0.0546
Cubic	Male	ZP	0.0091	0.0102
		ZPT	0.0125	0.0188
		UDP	0.0019	0.0044
	Female	ZP	0.0229	0.0136
		ZPT	0.0163	0.0173
		UDP	0.0015	0.0038

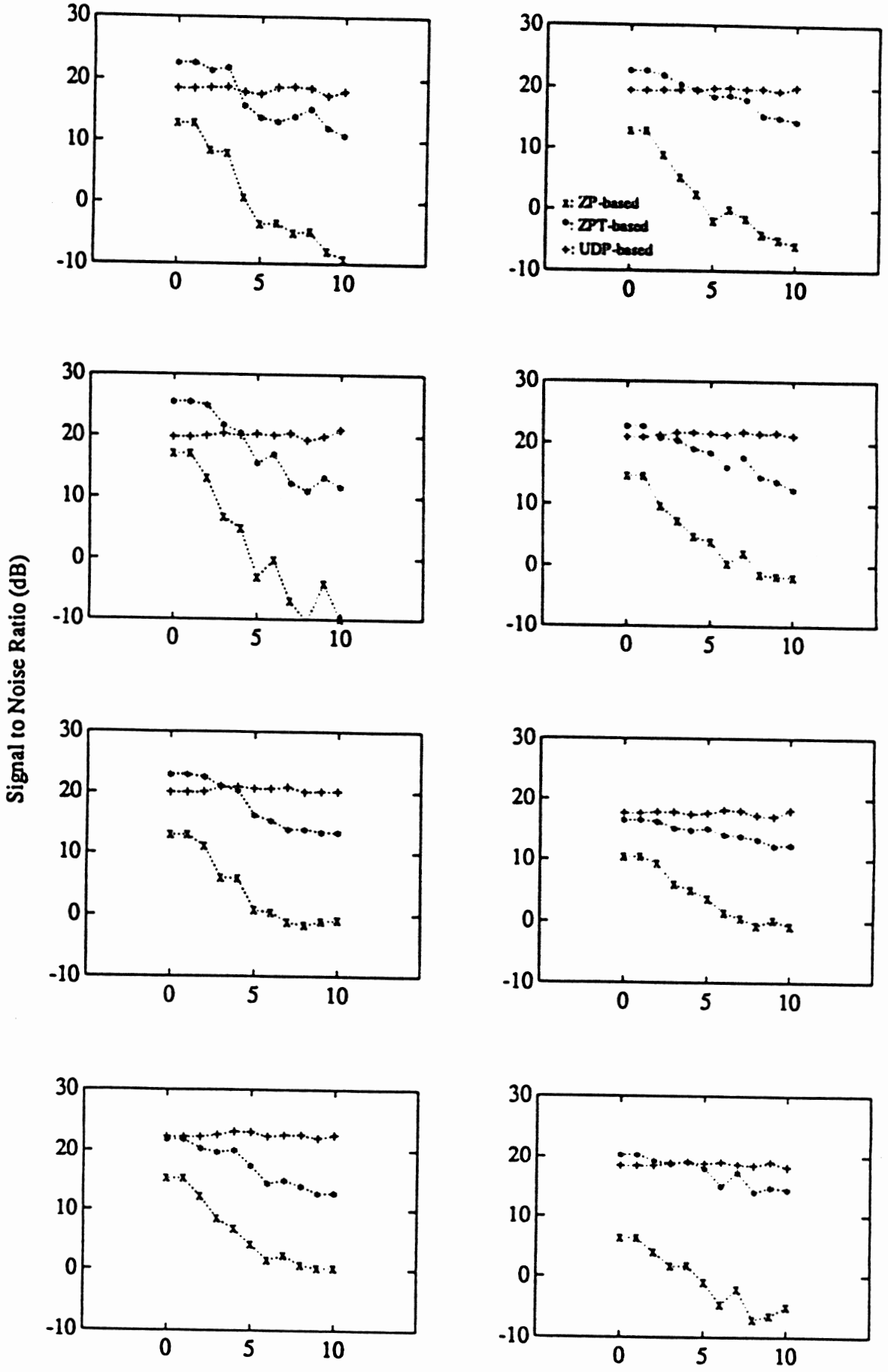
TABLE II. Results of Analysis of Variance

Parameter	Effect	DF	F Value	Pr. > F
Intercept	Gender	1	1.16	0.2910
	Method	2	50.93	0.0001
	GMI	2	1.90	0.1678
Linear	Gender	1	0.05	0.8313
	Method	2	28.93	0.0001
	GMI	2	0.36	0.6982
Quadratic	Gender	1	2.37	0.1347
	Method	2	4.34	0.0228
	GMI	2	0.88	0.4272
Cubic	Gender	1	3.40	0.0758
	Method	2	8.49	0.0013
	GMI	2	1.85	0.1753

GMI — Gender and Method Interaction


Signal to Noise Ratio (dB)
Maximum Error in Period Determination (Sample)
x: ZP-based
•: ZPT-based
+: UDP-based

(a)


Signal to Noise Ratio (dB)
Maximum Error in Period Determination (Sample)
x: ZP-based
•: ZPT-based
+: UDP-based
30
20
10
0
-10
0
5
10

(b)

Comparing reliability of perceptual and acoustic measures of voice[1]

C. Rose Rabinov, Jody Kreiman, and Bruce R. Gerratt
Division of Head and Neck Surgery,
UCLA School of Medicine
and
VA Medical Center, West Los Angeles

Acoustic analysis is often favored over perceptual evaluation of pathologic voice because it is considered objective, and thus reliable. "Subjective" ratings of voice quality are not highly regarded as either clinical or research tools, because of problems with intra- and interjudge reliability (e.g. Ludlow, 1981; Cullinan et al., 1963), because they are considered to lack objectivity and do not require great technical sophistication (Weismer & Liss, 1991), and because there is no accepted set of perceptual scales used by clinicians (e.g. Jensen, 1965; Yumoto et al., 1982). In part because of these views, so-called "objective," non-perceptual measures for vocal assessment have received much more attention in voice research. The assumption seems to be that some day acoustic measures may function in the place of perceptual assessment, thus alleviating concerns about listener unreliability.

However, recent studies suggest that this traditional bias in favor of acoustic analyses of voice may be unwarranted. Perceptual data (Kreiman et al., 1993; Gerratt et al., 1993) indicate that much of the noise in listeners' ratings is in fact predictable, and thus potentially controllable. Further, a study comparing several systems for perturbation measurement (Bielamowicz et al., 1993) suggested that agreement among different systems may be worse than assumed. Bielamowicz et al. compared values of jitter and shimmer produced by C-Speech (ver. 3.1), Kay CSL, SoundScope (ver. 1.09), and by an interactive hand marking system[2] developed at the VA Medical Center in West Los Angeles, for 50 voices ranging from normal to severely pathologic. Results for jitter are summarized in Figure 1. Analysis packages varied in their level of overall agreement, with Pearson's r for pairs of algorithms ranging from .21 to .77. However, even systems whose jitter measurements were moderately correlated did not necessarily produce the

---

[1] This research was supported in part by NIDCD grant # DC 01797 and by VA Merit Review Funds. Address correspondence to Jody Kreiman, VA Medical Center, West Los Angeles, Audiology and Speech (126), Wilshire & Sawtelle Blvds., Los Angeles, CA 90073.

[2] In this system, a waveform landmark (positive peak, negative peak, or zero crossing) that could be identified reliably from cycle to cycle was selected by hand. Perturbation measures were calculated using linear or parabolic interpolation, as appropriate (Titze et al., 1987).

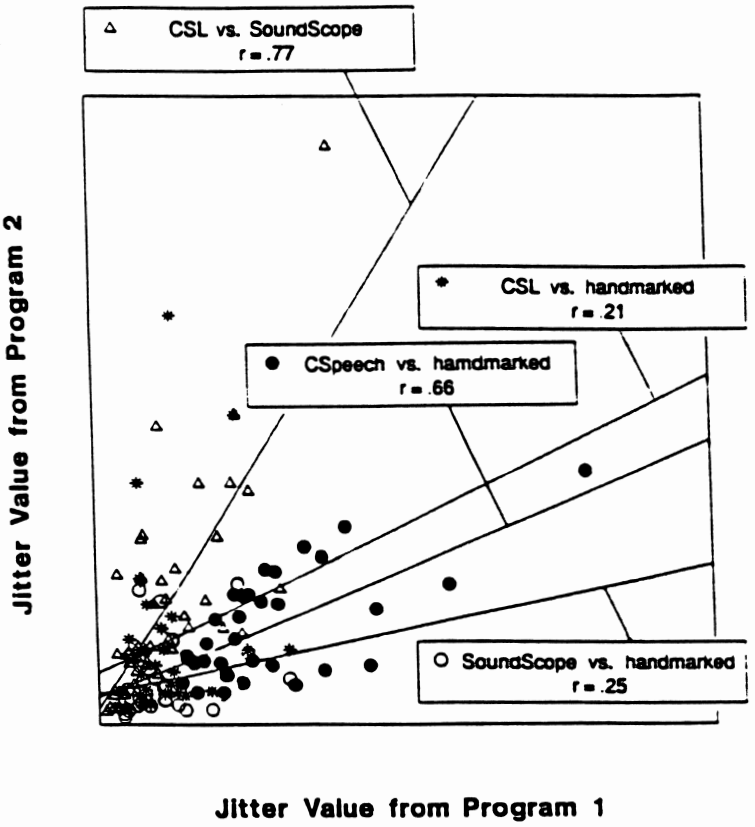


Figure 1. Comparison of jitter values produced by four analysis systems.

same numbers for a given voice. None of the lines in Figure 1 has a slope near 1, indicating that all packages systematically under- or over-estimated jitter relative to the others.

Two questions emerge from these findings. First, how do perceptual ratings of voice quality actually compare to acoustic measurements in reliability? Second, how similar are perceptual and acoustic analyses in their characteristics as measurement systems? That is, how similar are the patterns of agreement and disagreement among raters to those among analysis systems? To address these questions, we asked listeners to rate the roughness of the voice samples examined by Bielamowicz et al. (1993), and compared their ratings to the jitter measurements produced by the four analysis systems studied there.

METHOD

Listeners

Ten experienced listeners (otolaryngologists, speech pathologists, and phoneticians) participated in this experiment. Each had a minimum of two years' experience evaluating pathologic voice quality.

Stimuli

Fifty voices (29 male and 21 female) were selected from an existing library of samples. These voices were also used in the study by Bielamowicz et al. (1993) described above. Voices ranged from normal to severely disordered, with approximately the same number of voices at each of 5 severity levels.

Voice samples were originally obtained by asking speakers to sustain the vowel /a/. Utterances were low-pass filtered at 8 kHz, and a 2-second sample was digitized at 20 kHz from the middle of each utterance. Prior to the listening tests, digitized segments were normalized for peak voltage, and onsets and offsets were smoothed by 50 ms ramps to eliminate click artifacts.

Procedure

Listeners rated each voice twice, although they were not informed that any voices were repeated. Stimuli were rerandomized for each listener and were presented at a comfortable listening level in free field.

Listeners were tested individually in a sound-treated booth. Because jitter may be correlated with vocal roughness (e.g., Hillenbrand, 1988; Wendahl, 1966), they were asked to rate the roughness of each voice sample on a 7.5 cm visual analog scale, using whatever criteria for roughness they normally applied. The scale was displayed horizontally on a computer monitor, and had a resolution of 1 mm. Endpoints were labeled "not rough at all" and "extremely rough." Ten practice trials preceded the experimental session to familiarize listeners with the task.

RESULTS

Intrarater Agreement

Levels of test-retest agreement were acceptable for all listeners. Across listeners the correlation (Pearson's r) between the first and second ratings ranged from .75 to .90, with a mean of .83 (sd = .06). On the average, the first and second ratings differed by 9.8 mm (sd = 9.04).

Matched sample t-tests compared the first and second ratings of each voice, and indicated that ratings drifted significantly within a listening session. On the average, voices sounded significantly rougher at the second presentation than at the first ($t = -7.56$, $df = 499$, $p < .01$ one-tailed). Differences between the first and second ratings were also significant for 5 of the 10 individual raters ($p < .01$, adjusted for multiple comparisons). This drift is consistent with previous studies using unanchored rating protocols (Kreiman et al., 1993; Gerratt et al., 1993).

Of course, computer-based algorithms will always produce identical results under identical conditions. However, changes in analysis parameters within a given system did produce differences in results (Bielamowicz et al., 1993). Repeated independent analyses of 8

voices using the interactive hand-marking system produced mean jitter values within 0.05 ms in all cases (mean difference = 0.01 ms; sd = 0.02); percent jitter values were within 1% in all cases (mean difference = 0.20%; sd = 0.35%). (One voice was rejected as unmarkable in both analyses.) Mean jitter values produced by tokenized and untokenized analyses in C-Speech were correlated at .80; CSL analyses with tolerances of 1 and 20 ms were correlated at .47.

Interrater Reliability

Pairs of raters varied considerably in the extent to which their ratings agreed. Interrater agreement ranged from .32 - .90 (as measured by Pearson's r), with a mean of .71 (sd = .14), compared to a range of .21 to .77 for the different analysis systems (Figure 1). The intraclass correlation (ICC) was calculated using a mixed model ANOVA treating voices and listeners as random effects and presentations (first vs. second) as a fixed effect (model (2,1); e.g., Ebel, 1951; Shrout & Fleiss, 1979). This statistic reflects the overall cohesiveness of a group of raters, as compared to the pairwise comparisons above, and reflects the extent to which the present data might generalize to a new random sample of listeners. For the present data, the ICC = .64, consistent with the variability seen in the pairwise comparisons. Confidence intervals about the ICC were calculated using the formula in Shrout and Fleiss (1979). With 95% certainty, the true ICC value fell in the range $.54 < \rho < .75$.

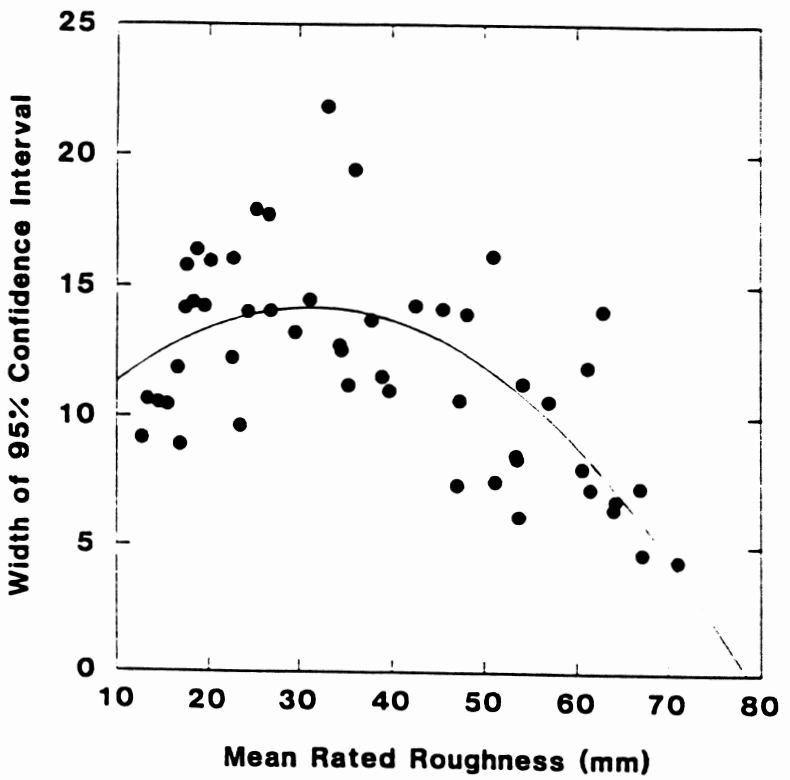


Figure 2. Variability in roughness ratings as a function of the mean rating.

Examination of patterns of agreement among pairs of raters suggested that subjects fell into two distinct "populations." One group included 7 raters; the other included 3. A one-way ANOVA showed that pairs of raters drawn from a single population agreed significantly better than pairs drawn from different populations ($F(1,43) = 52.30$, $p < .01$). Within a hypothetical population of raters, Pearson's r for pairs of raters ranged from .61 to .90, with a mean of .81 (sd = .07); across populations, r ranged from .32 to .81, with a mean of .60 (sd = .12).

Ratings of Individual Voices

Figure 2 shows the width of the 95% confidence interval (in mm) about the mean rating of each voice, plotted against the mean rating for that voice. The better the agreement among raters, the smaller the confidence interval. This figure shows the typical pattern (cf. Kreiman et al., 1993) of better agreement among raters (i.e., narrower confidence intervals) for voices at scale extremes, and worse agreement for moderately severe voices. The width of the confidence intervals ranged from 4.4 mm (± 2.2 mm) to 21.8 mm (i.e., ± 10.9 mm), for the 75 mm scale used here. In contrast, Figure 3 shows the 95% confidence intervals (in percent) around the mean of the percent jitter values produced by the different acoustic analysis systems. Uncertainty about measured jitter (indicated by larger confidence intervals) increases as a linear

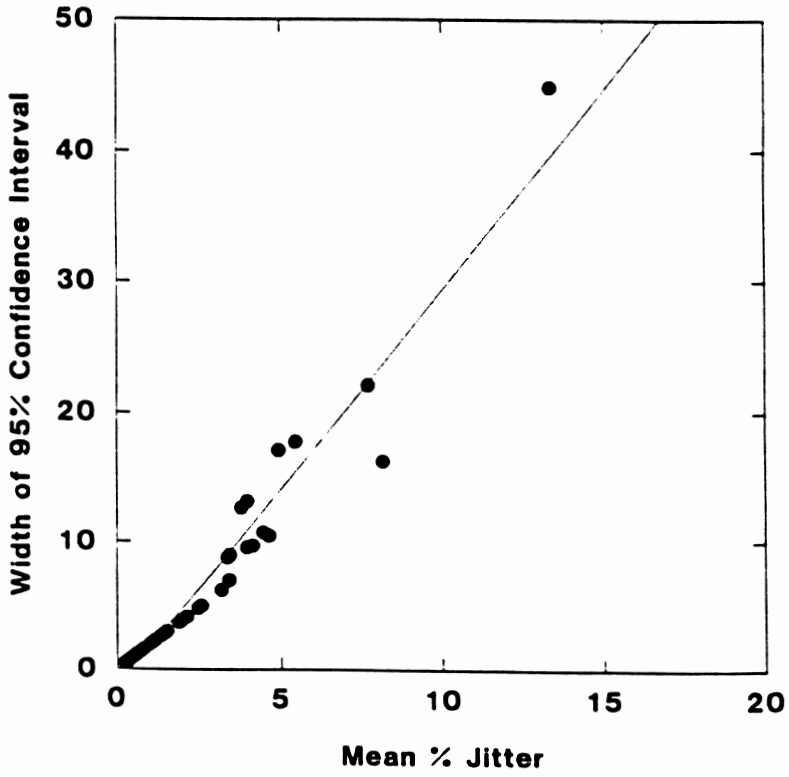


Figure 3. Variability in measured jitter as a function of the mean of values produced by four analysis systems.

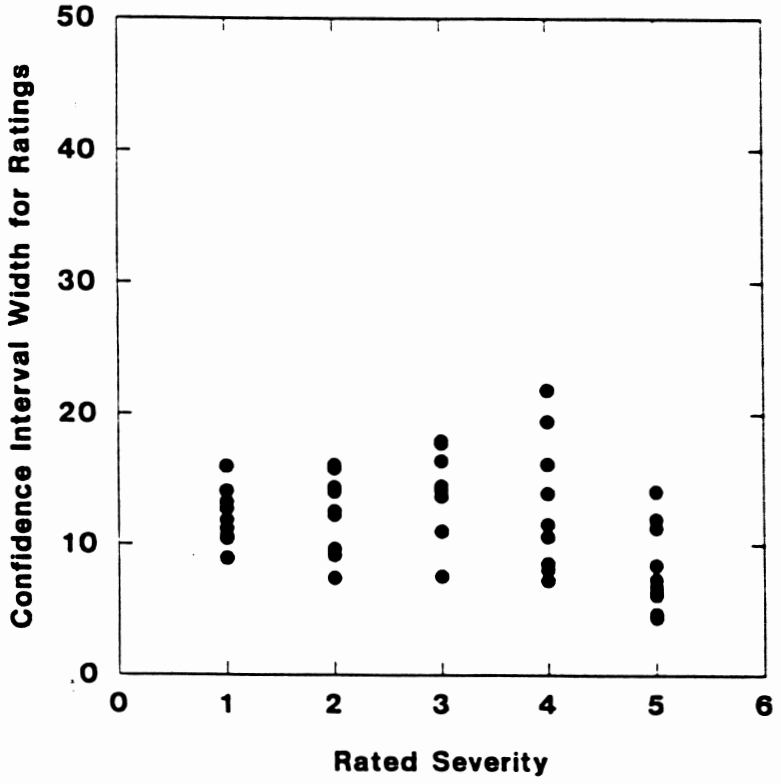


Figure 4. Variability in roughness ratings as a function of severity of pathology.

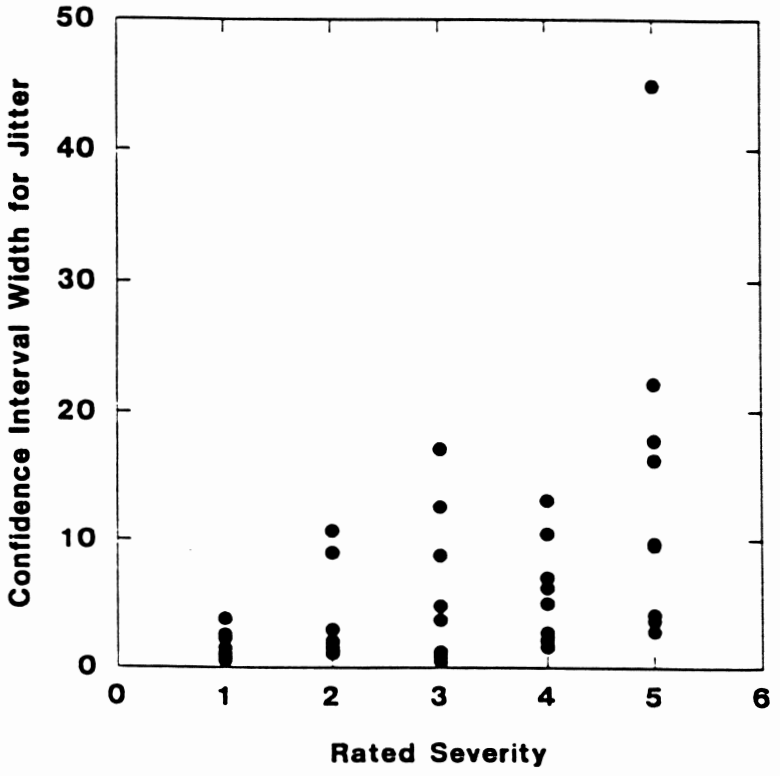


Figure 5. Variability in measured jitter as a function of severity of pathology.

function of the mean value ($F(1,47) = 878.11$, $p < .01$; $r^2 = .95$). Confidence interval width ranged from 0.48% to more than 44%.[3]

Figures 4 and 5 show how measurement uncertainty varies with severity of (perceived) pathology, for listeners and analysis systems respectively. For listeners, the range of variability in ratings increased slightly for voices with moderately severe pathology, again consistent with previous studies (Kreiman et al., 1993). That is, uncertainty about reliability is greatest for voices in the mid-range of pathology, and least for voices with mild or extremely severe pathology. In contrast, the "variability of the variability" associated with measured jitter increases with severity for the analysis packages, although packages apparently agreed about some voices at all severity levels.

Figure 6 shows the confidence intervals around mean voice ratings, plotted against the confidence intervals for mean jitter values. Data on both axes have been log transformed. This figure indicates that listeners tend to be most reliable when acoustic measures are most unreliable, and vice versa[4].

DISCUSSION

Levels of intrarater agreement in this study compare well to those in the literature (e.g., Kreiman et al., 1993; Gerratt et al., 1993), and represent good performance by experienced listeners. At least some test-retest disagreement is caused by systematic drift in ratings, which may be controllable by "anchored" paradigms using fixed comparison stimuli, as we have recently proposed (Gerratt et al., 1993). In one sense, intra-system reliability is not a serious issue for acoustic analyses, because computer-based algorithms will always produce identical results under identical conditions. However, changes in analysis parameters within a given system did produce differences in results (Bielamowicz et al., 1993). Recall that the correlation between listeners' first and second ratings of the voices ranged from .75 to .90. The correlation for analyses with different parameters within a given package ranged from .47 to .80. Thus across voices performance for even the worst listeners compared well with that of the most consistent analysis systems.

Across pairs of listeners, interrater reliability (measured by Pearson's r) ranged from .32 to .90; the ICC was .64. This compares well to Pearson's r for pairs of analysis systems, which ranged from .21 to .77. However, agreement levels among listeners improved greatly ($r = .61$ to .90) when listeners were compared only to others drawn from the same hypothetical "population of raters." The finding that listeners agreed and disagreed in groups is consistent with

---

[3] Mean jitter values produced by C-Speech were converted to percent jitter for this analysis.

[4] The regression is significant ($F(1,47) = 14.88$, $p < .01$); $r^2 = .24$.

multidimensional scaling studies of roughness (Kreiman et al., in press), which reported consistent differences in the strategies listeners used when judging roughness[5]. That study further demonstrated that differences in how listeners focus their attention on the different aspects of multidimensional perceptual qualities are a significant predictor of interrater agreement in voice quality ratings. Thus much of the variation in ratings within and across listeners may not in fact be noise, but may reflect the operation of consistent, predictable perceptual processes. A better understanding of these processes may lead to rating protocols which further enhance listener reliability.

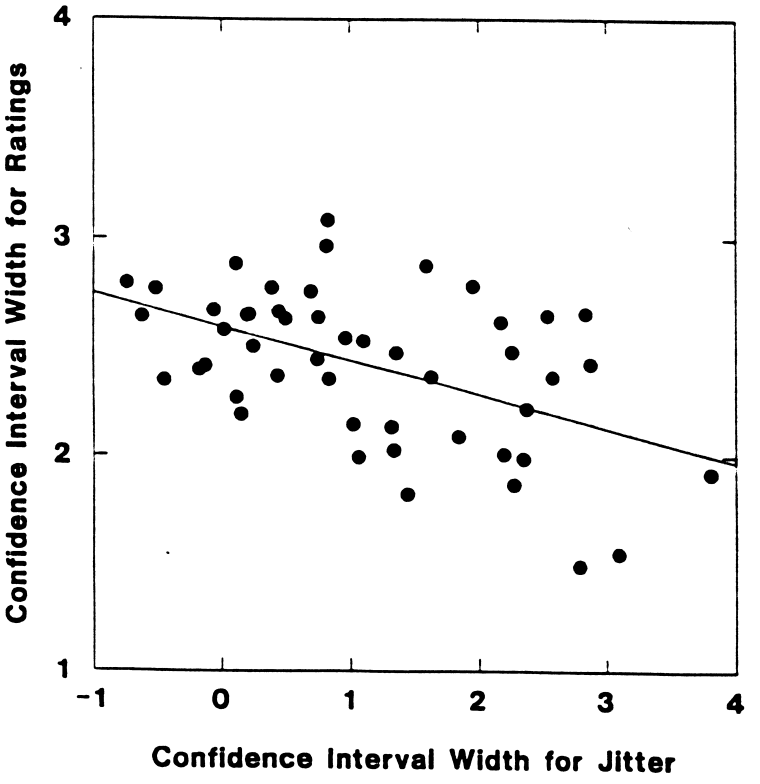


Figure 6. Variability in perceived roughness, vs. variability in measured jitter.

Interestingly, listeners and acoustic analysis algorithms differed in their properties as measurement systems. As Figure 6 showed, variability in jitter measures increased as rating reliability improved. Variability in ratings remained fairly constant across levels of severity, while variability in measured jitter increased dramatically with severity. These results suggest that acoustic measures have advantages over perceptual measures for discriminating among essentially normal voices. However, these advantages disappear once signals become irregular.

---

[5] In particular, listeners varied in how they handled breathy turbulent noise and tremor.

We therefore question the clinical assumption that acoustic measures may reasonably substitute for perceptual evaluation in the assessment of pathological vocal quality.

Our results suggest that measured jitter is a function of both signals and algorithms, much as perceptual measures are a function of both signals and listeners. While standardization of analysis techniques would solve the problem of disagreements among systems, a standard protocol will still represent a mapping between signals and measured values. The critical issue then becomes defining the "correct" algorithm, the choice of which must depend not only on technical considerations, but also on the purpose for which these measures are intended. As long as acoustic measures are used to detect or define pathology, to aid in diagnosis, to measure the extent of pathology, or to monitor treatment, they must reflect listeners' perceptions reasonably well. Standardization without attention to the characteristics of the application will result in measurements which are not useful.

In conclusion, listeners and analysis packages differ greatly in their measurement characteristics, but reliability is not a good reason for preferring acoustic to perceptual measures. Patterns of disagreement suggest that for clinical purposes perceptual measures are probably superior to current acoustic analysis systems. Standardization of acoustic measurement procedures without careful attention to all elements in the speech chain will not be fruitful.

REFERENCES

Bielamowicz, S., Kreiman, J., Gerratt, B.R., Dauer, M.S., and Berke, G.S. (1993). A comparison of voice analysis systems for perturbation measurement. Paper presented at the 125th Meeting of the Acoustical Society of America, Ottawa.

Cullinan, W.L., Prather, E.M., & Williams, D.E. (1963). Comparison of procedures for scaling severity of stuttering. *Journal of Speech and Hearing Research, 6,* 187-194.

Ebel, R. (1951). Estimation of the reliability of ratings. *Psychometrica, 16,* 407-424.

Gerratt, B.R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G.S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research, 36,* 14-20.

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America, 83,* 2361-2371.

Jensen, P.J. (1965). Adequacy of terminology for clinical judgment of voice quality deviation. *The Eye, Ear, Nose and Throat Monthly, 44* (December), 77-82.

Kreiman, J., Gerratt, B.R., & Berke, G.S. (in press). The multidimensional nature of pathologic vocal quality. To appear in *Journal of the Acoustical Society of America.*

Kreiman, J., Gerratt, B.R., Kempster, G., Erman, A., & Berke, G.S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21-40.

Ludlow, C. (1981). Research needs for the assessment of phonatory function. *ASHA Reports, 11,* 3-8.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Titze, I., Horii, Y., & Scherer, R.C. (1987). Some technical considerations in voice perturbation measurements. *Journal of Speech and Hearing Research, 30,* 252-260.

Weismer, G., & Liss, J. (1991). Reductionism is a dead-end in speech research: Perspectives on a new direction. In C. Moore, K. Yorkston, & D. Beukelman (Eds.), *Dysarthria and apraxia of speech: Perspectives on management* (pp. 15-27). Baltimore: Brookes.

Wendahl, R. (1966). Laryngeal analog synthesis of jitter and shimmer: Auditory parameters of harshness. *Folia Phoniatrica, 18,* 98-108.

Yumoto, E., Gould, W.J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America, 71,* 1544-1550.

The utility of acoustic measures of voice quality[1]

Bruce R. Gerratt and Jody Kreiman
VA Medical Center, West Los Angeles
&
Division of Head & Neck Surgery
UCLA School of Medicine

Acoustic analysis is becoming the preferred means of documenting normal and abnormal vocal qualities. These measures have long been popular in research applications, and the availability of off-the-shelf, automated programs now permits researchers and clinicians to generate acoustic values in almost real time fashion. Thus, use of these measures is becoming increasingly common in the clinic. Reflecting the increasing popularity and availability of acoustic measures, speech scientists and clinicians have focused much energy on technical and theoretical aspects of measuring vocal jitter and other aspects of vocal quality. We are here today, a little over 30 years after Lieberman's (1963) original paper on jitter, finally talking about ways to standardize these measures.

Although standardization is a worthy and noble goal, discussion of a measure's utility should be a prerequisite to the investment of more resources and effort. The utility of acoustic measures is occasionally questioned in the literature, particularly with respect to analyses of pathological voice (e.g., Hillenbrand, 1987), yet very little serious discussion has taken place in the scores of papers that have appeared in recent years. One notable exception is Catford, who argued in 1977 that the study of the acoustic signal without direct regard for the underlying physiology or its perception by a listener is without much purpose. An analogy to this argument is the study of the ink used in writing. Normally, the ink is useful only to the extent that its brightness, color, texture, and pattern conveys information from the writer to the reader. The careful investigation of these ink qualities in themselves is not usually informative about either the writer or the reader. Similarly, acoustic measures may shed light on physiologic or aerodynamic processes in speech; however, direct measures of these processes provide much better information, and are widely available.

What *are* the theoretical significance and practical uses of acoustic measures of vocal quality? This paper will focus on vocal perturbation, because of its great popularity for both

---

[1] This research was supported in part by NIDCD grant #DC01797, and by VA research funds.
Address correspondence to Bruce R. Gerratt, Audiology and Speech Pathology (126), VA Medical Center, West Los Angeles, Wilshire & Sawtelle Blvds., Los Angeles, CA 90073.

voice researchers and clinicians and because knowledge of the vocal period is a requirement for a number of other popular measures of voice. However, to varying degrees, our concerns apply to many other acoustic measures of voice quality. We will argue that perturbation measures have never been shown to correlate well with perceived vocal quality, have never been convincingly demonstrated to distinguish among pathological diagnoses, and in fact do not even consistently differentiate normal from pathological signals. Further, making these measures for voices that deviate from normal periodicity is technically difficult, if not impossible. In fact, the logic of measuring periodic deviation breaks down as the voices increasingly deviate from periodicity.

Problems Correlating Perturbation Measures and Physiology

Because perturbation of the signal can arise from a great number of sources, a particular jitter value (for example) can be accounted for by perturbation in muscular innervation, by secretions on the vocal folds, by mass deviations on the vocal folds, by tension asymmetry of the vocal folds, by randomness of flow through the glottis, by laryngeal tremor, by irregularities in source-vocal tract interactions through unstable articulatory configurations, or by almost any other deviation of laryngeal function (e.g. Titze, Horii, Scherer, 1987). In this way, jitter provides a very poor description of the actual laryngeal behavior which caused the perturbation observed in the acoustic signal. Thus, it is not surprising that researchers have found that jitter does not differentiate well among diagnostic categories (Hirano, 1989; Hirano, et al., 1988) nor even consistently separate pathologic from normal voices (Ludlow et al., 1987; Klingholz & Martin, 1983; Hecker & Kreul, 1971). Some researchers (LaBlance & Maves, 1992; Hirano et al., 1988) have argued that these measures are useful for documenting improvement following surgical treatment for voice disorders. However, in these studies perturbation values for both the pre- and post-treatment groups typically overlap with normal values, so it is difficult to know what such changes in jitter values actually signify.

Problems Correlating Perturbation Measures and Listener Perceptions

Despite the poor correlation of these acoustic measures to the underlying physiology, their relationship to listeners' perception of voice quality have traditionally been of great interest to researchers, for several reasons. First, the relationships among the various levels within the speech chain has its own inherent interest. Second, the historical difficulty in understanding and predicting quality perception has led researchers to seek a technology-based substitute for perceptual ratings.

However, understanding the relationship between acoustic and perceptual measures has proven difficult. Voices that are quite similar in quality can have quite different perturbation

measures, and voices that are quite different in quality can have perturbation levels which are similar.

In fact, studies examining the correlation between perturbation measures and perceptual qualities in disordered and normal speakers have had consistently negative results (e.g. Arends et al., 1990; Eskenazi et al., 1990; Heiberger & Horii, 1982; see Ludlow et al., 1987 for review). Studies using synthetic stimuli or using imitations of pathological qualities produced by normal speakers reported higher correlations (e.g. Coleman & Wendahl, 1967; Hillenbrand, 1988; Wendahl 1963, 1966). These better results are probably explained by the fact that the stimuli used vary primarily in only one dimension.

In contrast, pathological voices are perceptually complex, with many vocal qualities co-occuring and interacting. Studies using multivariate techniques and pathologic speakers have reported better correlations between perturbation measures and perceptual dimensions in multidimensional contexts (e.g. Kempster et al., 1991; Kreiman et al., in press; Eskanazi et al., 1990). Presumably, these improved correlations reflect the use of more appropriate perceptual models.

Thus, it appears that there is not a simple one-to-one correspondence between one perturbation measure and one perceptual quality. Traditional approaches seeking such associations imply far too simple a model of quality perception. Even traditional "qualities" such as breathiness and roughness may in fact be multidimensional, as we have recently argued. For example, voices described as breathy can include a diverse number of qualities which should not be funnelled into one perceptual construct. In fact, we found that a large source of listener variability (and an associated reduction in reliability) is that when listeners rate a voice on a perceptual scale, they pay attention to different dimensions of that quality (Kreiman et al., in press).

A further consideration is the dependence of quality perception on factors that are external to the voice itself. For example, the context within which a voice is judged has been shown to systematically affect the rating it receives (Gerratt, et al., 1993) Listeners' experience and perceptual habits also affect their perceptions (Kreiman et al., 1992; Kreiman et al., in press). In a study comparing listener groups, naive and expert listeners differed substantially in the perceptual strategies used to judge pathologic vocal quality; naive listeners primarily attended to F0 and deviation from normal, while experts used more complex, idiosyncratic perceptual strategies. In another study, differences between expert listeners in their judgments of roughness ranged from extreme (e.g., using unrelated perceptual strategies) to subtle (for example, using categorical vs. continuous dimensions for the same perceptual features). Better perceptual models may lead to voice judgment protocols which resist the effects of such variables. However, at present, it is naive to expect straightforward relationships between

acoustic signals and qualities, because acoustic signals do not provide information about these listener- and protocol-dependent effects.

The Paradox of Measuring the Unmeasurable

Another problem is the inherent difficulty in measuring perturbation in signals that deviate significantly from periodicity. This presents something of a paradox: As the phenomenon of interest (departure from periodicity) increases, confidence in determining periodicity (the essence of the measure) decreases. Several common voice types present particular difficulties (Figures 1-3). For some signals, periodicity cannot be defined with regard to the traditional concept of a single fundamental frequency. Other signals are simply too aperiodic for periods to be determined reliably. The greatest challenge here is making the decision of when to abandon the measurement result as invalid. At present, there are no accepted methods for making this decision.

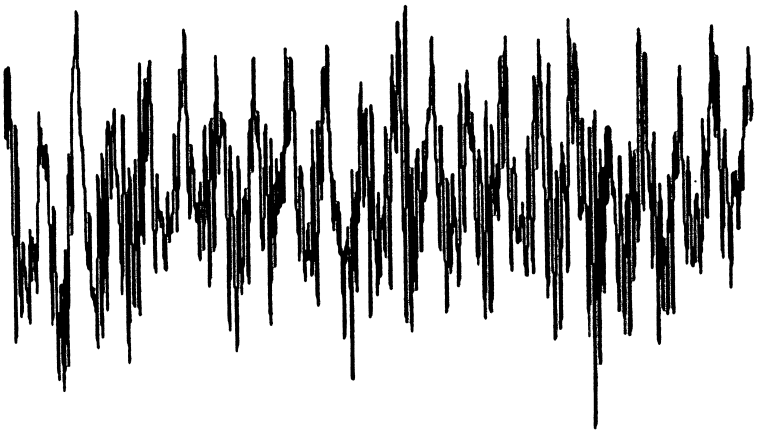


Figure 1. Acoustic waveform from a severely pathological voice: This 120 msec signal taken from a sustained production of /a/ corresponds to a severely breathy, moderately rough vocal quality, produced by a 39 year old man with chronic unilateral vocal fold paralysis after several attempts at Teflon augmentation.

Importantly, supraperiodic voice signals are fairly common among pathological and normal speakers. Klatt and Klatt (1988) reported that bicyclicity occurred in more than 25% of the utterances they examined. Both human operators and machine algorithms have great difficulty in measuring periodicity in these signals. The result is great disagreement among programs and among people. Thus, the practical problem of actually making the measurement in many of the voices which researchers and clinicians want to study also seriously reduces the utility of the measures.

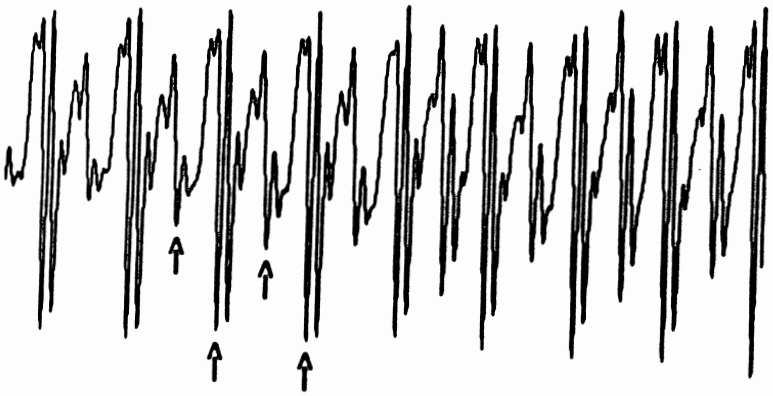


Figure 2. Bicyclic phonation (F0/2 subharmonic). There is noticeable period doubling in this 120 msec signal of a sustained /a/ from a 79 year old woman with vocal hyperfunction. Boundaries of vocal periods are marked by arrows. The corresponding auditory impression is a buzzing, mechanical, rough quality.

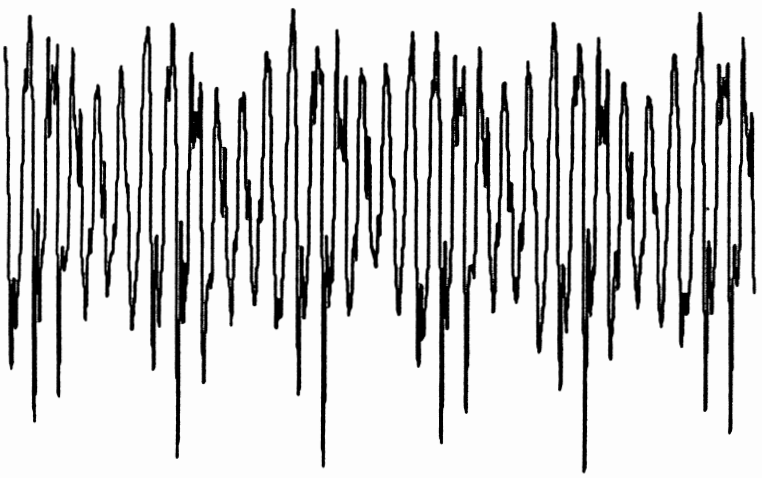


Figure 3. Diplophonic phonation (biphonation) - a high frequency wave (260 Hz) modulated by one of much lower frequency (44 Hz): This 120 msec signal represents a sustained /a/ produced by a 29 year old woman with a unilateral vocal fold paralysis. The corresponding auditory impression is rough, pulsed, and quite complex, with ambiguous pitch.

Conclusion

We have argued that categorizing and describing the acoustic signal is far less important than understanding its relationship to the other levels within the speech chain. However, we have presented some discouraging arguments regarding the possibility of relating acoustic measures to these other levels. Problems in correlating these measures to physiology appear unsolvable at present, an observation which has been made repeatedly by many. Significant theoretical and practical problems exist in relating acoustic measures to vocal quality perception,

although ultimately these may someday be alleviated by developing better perceptual models which include careful attention to interactions among the signal, the listener, and the listening task. Finally, periodicity is difficult to define for many voices, and some point exists beyond which vocal cycles cannot be identified reliably. Measuring jitter makes little sense for this category of voice. However, it is unclear at what point periodicity truly disappears, and it is unclear how such a point might be defined consistently. The limits of periodicity need better definition so that a user can know the level of confidence associated with a measurement result. Until these concerns regarding measurement utility are fully addressed, standardization of measures based upon vocal periodicity may proceed to an uncertain goal.

REFERENCES

Arends, N., Povel, D-J., van Os, E., and Speth, L. (1990). Predicting voice quality of deaf speakers on the basis of glottal characteristics. *Journal of Speech and Hearing Research*, 33, 116-122.

Catford, J. C. (1977) *Fundamental Problems in Phonetics* (Indiana University Press, Bloomington).

Coleman, F. and Wendahl,R. W. (1967). Vocal roughness and stimulus duration. *Speech Monographs*, 34, 85-92.

Eskenazi, L., Childers, D. G., and Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research*, 33, 298-306.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., and Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research*, 36, 14-20.

Hecker, M. H. L., and Kreul, E. J. (1971). Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency. *Journal of the Acoustical Society of America*, 49, 1275-1282.

Heiberger, V. L., and Horii, Y. (1982). Jitter and shimmer in sustained phonation. In N. J. Lass (editor), *Speech and Language: Advances in Basic Research and Practice, Vol. 7* (Academic Press, New York), pp. 299-332.

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

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America*, 83, 2361-2371.

Hirano, M. (1989). Objective evaluation of the human voice: Clinical aspects. *Folia Phoniatrica*, 41, 89-144.

Hirano, M., Hibi, S., Yoshida, T., Hirade, Y., Kasuya, H., and Kikuchi, Y. (1988). Acoustic analysis of pathological voice. *Acta Otolaryngologica (Stockholm)*, 105, 432-438.

Kempster, G. B., Kistler, D. J., and Hillenbrand, J. (1991). Multidimensional scaling analysis of dysphonia in two speaker groups. *Journal of Speech and Hearing Research*, 34, 53443-543.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, 820-857.

Klingholz, F. and Martin, F. (1983). Speech wave aperiodicities at sustained phonation in functional dysphonia. *Folia Phoniatrica*, 35, 322-327.

Kreiman, J., Gerratt, B. and Berke, G.S. (in press). The Multidimensional Nature of Pathologic Voice Quality. To appear in *The Journal of the Acoustical Society of America.*

Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research*, 35, 512-520.

LaBlance, G. R. and Maves, M. D. (1992). Acoustic characteristics of post-thyroplasty patients. *Otolaryngology - Head & Neck Surgery*, 107, 558-563.

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.

Ludlow, C., Bassich, C., Connor, N. P., Coulter, D. C., and Lee, Y. J. (1987). The validity of using phonatory jitter and shimmer to detect laryngeal pathology. In T. Baer, C. Sasaki, and K. Harris (editors), *Laryngeal Function in Phonation and Respiration* (College Hill, Boston), pp. 492-508.

Titze, I. R., Horii, Y., and Scherer, R. C. (1987). Some technical considerations in voice perturbation measurements. *Journal of Speech and Hearing Research*, 30, 252-260.

Wendahl, R. W. (1963). Laryngeal analog synthesis of harsh voice quality. *Folia Phoniatrica*, 15, 241.

Wendahl, R. W. (1966). Some parameters of auditory roughness. *Folia Phoniatrica*, 18, 26-32.

Establishment of Normal Limits for Speech Characteristics

Jon H. Lemke, Ph.D.
Division of Biostatistics
Department of Preventive Medicine and Environmental Health
University of Iowa, Iowa City, Iowa 52242

Hani M. Samawi, Ph.D.
Department of Statistics
Youmouk University
Irbid, Jordan

Introduction

Clinicians are often disenchanted by the possible diagnostic value of measures of voice and speech for various medical disorders associated with voice and speech. However, before the diagnostic value of these measures can be evaluated, we must establish normal ranges for these measures under standard conditions and across a variety of healthy subpopulations. If the percentiles are to be estimated by the crude maximum likelihood estimates, random sample sizes to establish normal ranges within acceptable limits usually are require to be 400 to 2000 subjects per subpopulation. If one employs resampling methods, one can substantially reduce necessary sample sizes. This translates to large savings in cost and time, making the establishment of normal limits more feasible than ever before.

Measures of voice perturbation are typically skewed and bounded below by 0.0, such as jitter, shimmer, the harmonic-to-noise ratio and the coefficients of variation for frequency and amplitude. For measures such as these one is typically required to sample at least 800 subjects to establish a normal range. Focusing on these four measures of voice perturbation, we will review the concept of normal range, present the data to be used for our examples, highlight resampling options and present antithetic resampling results for examples from a sample of 47 disease-free males (Ramig and Ringel, 1983).

This work has been supported in part by the National Center for Voice and Speech, NCVS, with support form the National Institute on Deafness and Other Communication Disorders (Grant: P60 DC00976) and in part by the University of Iowa College of Medicine Departments of Preventive Medicine and Environmental Health, and Otolaryngology.

Background

Concept of Normal Range

By definition, the normal range of a continuous variable contains 95% of all disease-free individuals in a population. Recognize that abnormal and disease-free are not synonymous. By definition, 5% of the disease-free people must have abnormal values. When a person has a characteristic in the normal range, one is not guaranteed to not have associated medical problems. Changes within the normal range may be pathologic and indicative of a medical problem. The diagnostic value of a variable depends upon the distribution of a variable for people with a certain medical condition and the prevalence of the medical condition in the population that gets referred for evaluation. If everyone in the general population gets tested and the prevalence in the general population is low, the predictive value positive will be minimal even for sensitivities in the range of 99%.

Different populations may have different normal ranges. Males and females certainly have different normal ranges for fundamental frequencies and one expects them to have different normal ranges for other variables as well. Especially for variables which measure maximum performance speech tasks, one will find the normal ranges will change with age (Ramig and Ringel, 1983). Age-related changes have been reported for fundamental frequency, maximum phonation range and average jitter by Mysak (1959), Endres, Bambach and Flosser (1971), Segre (1971), Hollien and Shipp (1972), and Wilcox and Horii (1980). For acoustic voice measures, the normal range typically becomes wider as the upper limit changes more rapidly then the lower limit reflecting an increase in intersubject variability with age. If the measure is a perturbation measure bounded below by 0.0, the lower limit should be set at 0.0 and not change at all. The upper limit should be the 95th percentile (P_{95}) and may be expected to increase with age. Whereas, other measures such as maximum duration of sustained phonation or maximum phonation range should have a normal range bounded below by $P_{2.5}$ and above by $P_{97.5}$. For these factors they can be expected to decrease with age. There are ethnic differences in some normal ranges.

It may not be desirable to be in or stay in the normal range. Just as basketball players prefer to be abnormally tall, singers and orators value their voices with abnormally high or low pitch or abnormally wide pitch range. Olympic records are set by people with abnormal abilities. If a normal range is changing with age and you were normal, maintaining the youthful level may be just fine even when it is abnormal.

The normal range is not to be confused with the normal distribution, even though the normal range is often explained using the normal distribution. In the normal distribution case one typically wants the middle 95% disease-free individuals defined as normal, leaving 2.5% on each side. In general one can split the 5% disproportionately provided it makes sense pathologically, as long as you capture 95% of the disease-free individuals. Perturbation measures are skewed. The whole idea of telling somebody that they are abnormal with a jitter of 0.1% or with any other perturbation measure very close to 0 doesn't have any diagnostic value when the measure is associated with a disorder. You only want to consider a person abnormal if the value is excessively high. For perturbation measures we always recommend using the 95th percentile (P_{95}) as the upper the normal limit and 0.0 as your lower limit.

Distributions of Perturbation Measures

Ramig and Ringel (1983) present perturbation measure data conditional upon 47 male subjects' exercise level. Their objective was to contrast healthy men who were active vs. inactive. The subject pool was partitioned into thirds by activity level with the middle 1/3 excluded. Thus, the data does not represent a random sample of men and the presented estimates should tend to overestimate the upper limits of the normal ranges.

The perturbation measures of jitter and shimmer are defined by

$$\frac{100\sum_{i=2}^{N}|x_i - x_{i-1}|}{(N-1)\bar{x}}, \text{ where N is the number of consecutive cycles.}$$

For jitter the x_i is the frequency of the ith cycle and for shimmer the x_i is the amplitude of the ith cycle. The coefficients of variation for frequency and amplitude are determined by

$$\frac{100}{\bar{x}}\sqrt{\frac{\sum_{i-1}^{N}(x_i - \bar{x})^2}{N}}.$$

Jitter and shimmer are means of absolute values and thus always nonnegative. Coefficients of variation in frequency and amplitude are nonnegative with very skewed distributions and are typically analyzed inappropriately using normal distribution theory. Similarly, the harmonic-to-noise ratio must be nonnegative and is very skewed.

We present the data from Ramig and Ringel (1983) using stem-and-leaf plots, which retain all the data unlike histograms where the actual values of the data are lost. The data is also ordered and reflects the shape of the distribution. Consider Figure 1, where we present the jitter data. The stem is the middle column of numbers and the leaves are to the right. There are 47 leaves for the 47 subjects. For jitter the leaf unit is to the hundredth of a percent. The first row 6 | 2 | 455789 represents 0.24%, 0.25%, 0.25%, 0.27%, 0.28% and 0.29%, the six lowest jitters (leaves) with the 6 on the left representing the cumulative frequency through the first level. The totals on the left are cumulative from both the minimum value on the top and the maximum value from the bottom. The (9) on the left of the third row indicates the number of leaves on the branch with the median value ($\hat{P}_{50}$) of 0.43%, which is easily identified by using the cumulative counts either from the top or the bottom. The data ranges from a minimum of 0.24% to a maximum of 1.13%, represented by the last leaf at the bottom of the figure. The crude maximum likelihood estimate ($\hat{P}_{95}$) of P_{95} is 0.98.

6	2	455789
18	3	022223556666
(9)	4	111223789
20	5	1249
16	6	23789
11	7	2347
7	8	78
5	9	478
2	10	6
1	11	3

Figure 1: Stem-and-leaf plot of jitter (leaf unit = 0.010%)

The stem-and-leaf plots for shimmer, harmonic-to-noise ratio, and coefficients of variation for amplitude and frequency appear in Figures 2-5, respectively. The shimmer values range from 0.8% through 7.2% with a median value $\hat{P}_{50}$ of 1.7% and $\hat{P}_{95}$ is 6.1%. The harmonic-to-noise ratios vary from 14.7 to 25.6 with a $\hat{P}_{50}$ of 20.6 and $\hat{P}_{95}$ equal to 24.6. The coefficient of amplitude ranges from 2.0 to 15.3 with $\hat{P}_{50} = 6.2$ and $\hat{P}_{95} = 11.3$; and the coefficient of frequency ranges from 0.52 to 2.16 with $\hat{P}_{50} = 0.94$ and $\hat{P}_{95} = 1.61$.

4	0	8999
16	1	000012222234
(10)	1	5556667789
21	2	0222333444
11	2	689
8	3	01
6	3	
6	4	00
4	4	5
3	5	
3	5	
3	6	12
1	6	
1	7	2

Figure 2: Stem-and-leaf of shimmer
(leaf unit = 0.10%)

1	14	7
3	15	59
4	16	3
4	17	
10	18	037889
18	19	12246668
(10)	20	2244466999
19	21	234458999
10	22	23557
5	23	2
4	24	56
2	25	46

Figure 3: Stem-and-leaf for the
harmonic-to-noise ratio
(leaf unit = 0.10)

2	2	01
6	3	0388
11	4	11258
21	5	2234678899
(9)	6	112222236
17	7	124
14	8	0244599
7	9	36
5	10	8
4	11	038
1	12	
1	13	
1	14	
1	15	3

Figure 4: Stem-and-leaf for the
coefficient of variation of amplitude
(leaf unit = 0.10)

3	5	226
6	6	158
11	7	02259
19	8	01135999
(8)	9	12344569
20	10	23344456
12	11	
12	12	0366788
5	13	0
4	14	2
3	15	
3	16	18
1	17	
1	18	
1	19	
1	20	
1	21	6

Figure 5: Stem-and-leaf for the coefficient
of variation of frequency
(leaf unit = 0.010)

Resampling Methods

We will briefly review the bootstrap (or uniform resampling) and antithetic resampling methods. The bootstrap is a uniform resampling of your data where you randomly sample the original data with replacement. The technique is very useful whenever the sampling distribution of an estimator is unknown. (Efron, 1990) From this group of 47 observations, randomly sample 47 observations with replacement. Some observations will be selected more than once and some will not be selected at all. One random sample doesn't provide you with an estimate of the standard deviation of the sampling distribution of the estimator. Thus, you take, say, 100 random samples with replacement from your original data. The distribution of the 100 estimates approximates the sampling distribution of the estimator and then confidence intervals can be established for the estimated parameter even when they may not be available theoretically.

To understand antithetic resampling, let $x_{[1]}$, $x_{[2]}$, $x_{[3]}$, ... $x_{[n]}$ represent the original n observations sorted from the minimum $x_{[1]}$ to the maximum $x_{[n]}$. If there are ties, they are replicated as in the stem-and-leaf figures. Each antithetic sample estimate begins with a uniform resampling just as in the bootstrap. Now, each sample is paired with another sample where the ranks are the reverse in the following sense: if $x_{[2]}$ is in the initial sample, then $x_{[n-1]}$ is in the antithetic pair. For n = 47, there are the same number of $x_{[46]}$'s in the paired sample as there are $x_{[2]}$'s in the first sample. The estimates from the antithetic paired samples are negatively correlated, since an overestimate of the parameter in one sample provides a tendency to underestimate the parameter in the other sample of the antithetic pair. The two estimates in each pair are averaged to obtain the estimate from the pair. Taking the estimates from 100 antithetic pairs, one obtains an approximation of the sampling distribution of the antithetic resampling estimator. Since the antithetic pair estimates are negatively correlated, one is much better off taking a random resampling of 100 antithetic pairs than one is taking 200 bootstrapped resamples.

We performed extensive Monte Carlo studies across a variety of symmetric and skewed distributions and sample sizes (64 and 100) to compare crude maximum likelihood estimates, bootstrap resampling, importance resampling and antithetic resampling, we found the antithetic resampling estimate was the least biased and had the smallest mean square error (MSE) about the actual value of P_{95} being estimated. The MSE of the estimates of P_{95} using antithetic resampling was at least 17% less than the crude estimates in all cases of underlying distributions and sample

sizes. To highlight the accuracy when estimating $P_{95} = 9.488$ from a Chi-square distribution with 4 degrees of freedom based upon 1000 simulations, the mean crude maximum likelihood estimates and mean antithetic resampling estimates were 8.928 and 9.344, respectively.

When the actual parametric form of the underlying distribution function is known, then solving for the maximum likelihood estimator is recommended over antithetic resampling. However, the benefit is small and if the underlying distribution is questionable or possibly a mixture of distributions, then we strongly recommend antithetic resampling or a form of importance sampling with appropriately chosen weights. (Kim-Anh and Hall (1991), Hall (1991) and Johns (1988)).

Results

Having taken 100 random resamples from each of the observed distributions of perturbation measures, we present the crude maximum likelihood estimates and antithetic resampling estimates of P_{95} in Table 1. In addition, we include an estimated standard error of $\hat{P}^*_{95}$ (the standard deviation of the antithetic estimates), the corresponding confidence intervals and an estimated relative efficiency. To couple variance estimates for the relative efficiency we divided the variance of the 200 unpaired bootstrap estimates by the variance of the 100 estimates from the antithetic pairs.

The estimates $\hat{P}^*_{95}$ and $\hat{P}_{95}$ of P_{95} varied depending upon the observed tails of the perturbation measure distributions. The greatest difference appeared with shimmer where the crude estimate is 15% greater than the antithetic estimate. Even though the estimates may not vary dramatically, note in the stem-and-leaf plots how dramatically one could vary the crude estimates by adding a single observation to the tail of the distribution.

The 95% confidence intervals for P_{95} are fairly wide with a sample size as small as 47. There is also a 0.090 probability with $n = 47$ that all observations were less than P_{95}. When this occurs the estimates are obviously underestimates. With a minimum random sample size of 100 this probability drops substantially to 0.0059. One should not attempt to establish normal limits with sample sizes of 47 even with antithetic resampling.

The minimum estimated relative efficiency is 2.2. Thus, one will have a savings of at least 55% in the number of subjects required to obtain an estimate of P_{95} for these perturbation measures. With $n = 47$ we have antithetic estimates which would have required sample sizes of at

least 103 (47*2.2) for crude estimates. If one would be required to random sample 800 disease-free individuals in a subpopulation, one now needs 364 individuals to obtain the same degree of accuracy. If one wants to estimate percentiles further from the median, the relative efficiency decreases as the negative correlations approach 0. If one wants to estimate percentiles closer to the median, the negative correlations approach -1 and the relative efficiency increases dramatically.

Table 1: Crude maximum likelihood estimates and antithetic resampling estimates of P_{95} for various perturbation measures using the 47 subjects in Ramig and Ringel (1983).

	Maximum Likelihood $\hat{P}_{95}$	Antithetic Estimate $\hat{P}^*_{95}$	$SE(\hat{P}^*_{95})$	95% Confidence Interval for P_{95}	Estimated Relative Efficiency
Jitter	0.98	0.9835	0.0462	(0.8930, 1.0741)	2.4
Shimmer	6.13	5.3369	0.7873	(3.7938, 6.8800)	2.2
Harmonic -to- Noise Ratio	24.63	24.431	0.623	(23.210, 25.652)	2.2
Coefficient of Variation of Amplitude	11.35	11.351	0.798	(9.787, 12.915)	2.3
Coefficient of Variation of Frequency	1.61	1.5418	0.1465	(1.2547, 1.8289)	2.2

The correlations of the antithetic pairs for jitter, shimmer and harmonic-to-noise ratio are -0.181, -0.074 and -0.076, respectively. For the coefficients of variation the correlations are -0.149 and -0.084 for amplitude and frequency, respectively. These negative correlations were crucial to obtain smaller variances through antithetic resampling than for crude estimates or bootstrap resampling estimates. The benefit becomes apparent when you consider Figures 6-7, where the resampled estimates for the coefficient of variation of amplitude are the means for the pairs in Figure 6 and the individual unpaired estimates in Figure 7. When paired, the standard deviation is 0.798; while unpaired, the standard deviation is 1.217.

	9	
1	9	3
1	9	
9	10	02222333
16	10	4455566
19	10	779
48	11	00000000000001111111122222222
(37)	11	44444444444444444444466666666666666
		6
15	11	888
12	12	22
10	12	
10	12	
10	13	122223333
1	13	6
	13	
	14	
	14	
	14	
	15	
	15	

Figure 6: Stem-and-leaf plot for the 100 antithetic pair estimates of the coefficient of variation of amplitude (leaf unit = 0.10)

6	9	000000
11	9	33333
21	9	6666666666
21	10	
21	10	
49	10	99999999999999999999999999
88	11	0000000000000000000000000000000000000
(59)	11	44444444444444444444444444444444444444444444444444444444444
53	11	88888888888888888888888888888888888888
12	12	
12	12	
12	12	
12	13	
12	13	
12	13	
12	14	
12	14	
12	14	
12	15	
12	15	333333333333

Figure 7: Stem-and-leaf plot for the unpaired 200 estimates of the coefficient of variation of amplitude (leaf unit = 0.10)

Discussion

The diagnostic value of any voice characteristic for a specific voice or speech disorder depends upon knowledge of the normal range of the characteristic for disease-free individuals, as well as the distribution of the variable for subjects with the disorder. Current dissatisfaction with the diagnostic value of many measures exists with a lack of knowledge of the normal range. Especially, for measures of perturbation there will be many conditions where there is substantial overlap between the perturbation measure distribution of the disease-free and those with the disorder, and there will be little diagnostic value if any. The diagnostic value varies both by voice or speech disorder and voice characteristic. With the advent of antithetic resampling one can practically establish normal limits for voice measures using substantially fewer disease-free subjects. Since one must expect the normal limits to vary by age, gender and ethnicity, the savings is magnified many times when establishing a comprehensive set of normal limits.

It is crucial that the set of voice measures that has diagnostic value for any specific disorder is as small as possible in order to eliminate the random possibility of diagnosing an excess number of disease-free individuals as having a voice disorder, that is having a low specificity. Thus, for each voice disorder one needs to restrict the number of voice characteristics considered to have diagnostic value to at most four or five. Given the correlation between many voice measures, the preferred sets should have measures which are orthogonal to each other.

If one's level of physical activity effects the perturbation measures assessed, then the estimates in Table 1 are likely overestimates of the true P_{95}'s. This would be the likely case if one should expect less than 5% the 24 males excluded from the Ramig and Ringel (1983) study to have values less than the estimated P_{95}'s. Since the Ramig and Ringel sample of 47 was not a random sample of disease-free subjects, the emphasis of this article is not on the estimated values but upon the antithetic resampling method to estimate the normal ranges..

In speech analysis subjects are at a premium, since it can be expensive and time consuming; thus we require resampling methods such as antithetic resampling to obtain good estimates of the parameters of interest. This is true for other measures of speech performance besides percentiles for disease-free subjects. This technique should be expanded to estimation within tokens and across tokens to properly evaluate individual patients.

References

Do, Kim-Anh and Hall, Peter (1991). On importance sampling for the bootstrap, Biometrika, 78(1): 161-167.

Efron, Bradley (1990). More efficient bootstrap computations, Journal of the American Statistical Association, 85(409): 79-89.

Endres, W., Bambach, W. and Flosser, G. (1971). Voice spectograms as a function of age, voice disguise and voice imitation. Journal of the Acoustical Society of America, 49:1842-1848.

Gelfand, Alan E. and Smith, Adrian F. M. (1990). Sampling-based approaches to calculating marginal densities, Journal of the American Statistical Association, 85(410): 398-409.

Hall, Peter (1991). Bahadur representations for uniform resampling and importance resampling, with applications to asymptotic relative efficiency, The Annals of Statistics, 19(2): 1062-1072.

Hammersley, I. M. and Handscomb, D. C. (1964). Monte Carlo Methods, John Wiley, New York.

Hinkley, D.V. and Shi, S. (1989). Importance sampling and the nested bootstrap, Biometrika, 76(3): 435-446.

Hollien, H. and Shipp, T. Speaking fundamental frequency and chronological age in males. Journal of Speech and Hearing Research, 15: 155-159.

Johns, M. Vernon. (1988). Importance sampling for bootstrap confidence intervals, Journal of the American Statistical Association, 83(403): 709-714.

Mysak, E.D. (1959). Pitch and duration characteristics of older males. Journal of Speech and Hearing Research, 2:46-54.

Ramig, L.A. and Ringel R.L. (1983). Effects of physiological aging on selected acoustic characteristic of voice. Journal of Speech and Hearing Research, 26: 22-30.

Segre, R. (1971). Senescence of the voice. Eye, Ear, Nose and Throat Monthly, 50:223-233.

Serfling, Robert J. (1980). Approximation Theorems of Mathematical Statistics, John Wiley & Sons, New York.

Wilcox, K.A. and Horii, Y. (1980). Age and changes in vocal jitter. Journal of Gerontology, 35: 194-198.

Jitter Measurements on Aging Voices

Edward P. Neuburg
Center for Communications Research
Thanet Road, Princeton, NJ 08540
e-mail: epn@ccr-p.ida.org

1. Introduction

As data for a study of changes in voice with age, Arthur House has two sets of recordings of three male subjects, the first made in 1960, when the subjects were all about 37 years old, and the second in 1990 when they were about 67. Certain findings about vowel duration and amplitude have already been reported. (House and Stevens 1993). I have been working with Dr. House on a continuation of the study, measuring other speech features such as intonation patterns and excitation characteristics. Here two aspects of this work are discussed: automatic location of pitch epochs, and estimation of jitter given these epochs.

(Note: the phrase "pitch period" is used in the literature ambiguously: sometimes it means *the length of time between glottal closures*, and sometimes it means *the segment of speech starting at one glottal closure and ending at the next*. Here I use "pitch period" for the first, and "pitch epoch" for the second.)

2. The Speech Material

There are three talkers, AH, JM, and KS. Each talker produced the same set of bisyllabic nonsense tokens, consisting of a vowel imbedded between identical consonants, and preceded by an unstressed /HAX/. Thus typical stimuli sound like "hubob", "hutat", "hugig". Each talker went through exactly the same set of utterances in his "old" recording as he had done in his "young" recording (but in a slightly different order).

Tokens were recorded on tape, then digitized at a 10kz sampling rate, 12 bits per sample. Hum and noise are negligible.

In the full study, there are 10 vowels and 23 consonants. For this paper, I selected 12 consonants and three vowels, /AA/, /IY/, and /UW/, and extracted from each token a sub-token consisting of the "middle half" of the vowel; that is, the signal starting 1/4 of the way into the vowel, and ending 3/4 of the way into the vowel (according to Dr. House's hand marking). Table 1 gives the total number of pitch epochs per vowel.

	Talker					
	young	old	young	old	young	old
Vowel	AH	AH	JM	JM	KS	KS
/AA/	172	170	169	177	135	133
/IY/	146	154	144	144	113	104
/UW/	165	170	146	143	113	107

Table 1: Number of pitch epochs per vowel, per talker, per age

3. On Finding Pitch Period

If we were observing the larynx, we might take as pitch period the time between occurrences of some well-defined event in the laryngeal cycle. However we are not observing the larynx, or even capturing the signal at the larynx; we have to make do with a poorly-understood transformation of that signal.

If pitch and amplitude are steady, and formants and excitation are not changing, intervals between identical events in successive epochs might be a usable approximation to the desired laryngeal intervals. An example would be the time of occurrence of the largest peak in the pitch epoch. We would also expect, in this case, that the interval at which the signal autocorrelation has a maximum will be the same as this interval between well-defined events; and indeed, peak-picking and autocorrelation, the two most widely-used methods for finding pitch period, usually yield very similar estimates. (See e.g. Titze and Liang 1993)

However, if any or all of these properties are changing, it's a whole new ball game. If pitch is changing, what exactly do we mean by "pitch period"? At the larynx, there is perhaps a physical event that occurs once per "cycle"; but in the speech signal, what part of two successive epochs should we use as benchmarks to measure interval? Or if we use autocorrelation, how should it be done? (Because the length and type of correlation certainly affect the location of its maximum.)

If formants are moving, or the excitation is changing, successive epochs look different, and there may be no event in two successive epochs to use for an interval measurement. If, in the speech signal, we could find the moment of a laryngeal event such as closure, we could use that as the benchmark; I know no way of doing this reliably, especially when the form of the epoch is changing.

The tokens in this study contain all these sources of variability. They are spoken with "list intonation", the style in which subjects produce every word or phrase in the list, with falling pitch. (The pitch may actually rise a little at the start.) Figure 1 illustrates this; you can see that the pitch periods at the end of the utterance are longer than those at the beginning. (You can also see that amplitude is decreasing, another feature of this speaking style.)

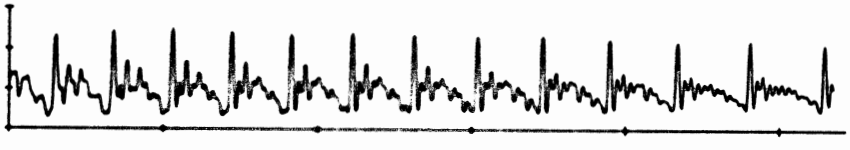


Figure 1: A typical token

Since the vowel is surrounded by consonants, the formants move, especially at tbe beginning and end of the vowel. In Figure 1 you can see how formant movement is causing differences in shape of successive pitch epochs. There are also some tokens with change in excitation, and it is these that give the most trouble to the pitch tracker and epoch finder described below.

If we needed to determine jitter absolutely (or clinically), these difficulties would be formidable. However what we want in our study is a comparison between jitter now and jitter 30 years ago, or jitter of AH and jitter of KS, say. And until and unless we look at jitter for vowels in their individual contexts, we have plenty of data; at least 100 pitch epochs for every talker/vowel. Thus (and perhaps this applies in the clinical case as well) consistency is more important than accuracy. We have so far considered four working definitions of jitter, all of which depend on a certain epoch-finding algorithm, which is described in the next Section. The measures of jitter are discussed in Section 5.

4. Finding the Pitch Epochs

Our algorithm for finding pitch epochs involves two kinds of speech analysis and two dynamic programs. It is complicated because in real speech, no single pitch-finding method (known to me) finds all individual pitch epochs reliably, without occasionally making an unacceptable error. Since in our study we have a great number of tokens to deal with, and cannot look at all pitch tracks for all tokens, and since it doesn't take many doubled or halved pitch periods to bias a statistical test, we need a pitch-epoch-finder that we can trust to operate without error on reasonably clean speech. The algorithm proceeds as follows:

1) Form the smoothed, half-wave-rectified LPC residue for the entire token;
2) From the residue, estimate pitch period every 100 samples;
3) Find all the peaks in the rectified LPC residue;
4) From peaks, select "pitch pulses" that divide the token into epochs;
5) At each pitch pulse, reestimate pitch period from the original speech.

4.1 The LPC Residue

The LPC residue is created by doing a 12-coefficient LPC at centisecond (100 sample) intervals, using 150 Hanning-windowed samples, and putting each 100 samples back through its local inverse LPC filter to obtain the current 100 samples of residue. The first line in Figure 2 shows a typical speech token, the second line shows its LPC residue.

The next step is a mild low-pass filtering of the LPC residue. The filter is a crude one: the output is the sum of the amplitudes of the last 5 samples, plus the sum of the last 6 samples, plus the sum of the last 7 samples.

In preparation for picking peaks, the residue is tested to determine whether the positive-going peaks or the negative-going ones are more prominent. The sum of squares of the positive values in the residue is compared with the sum of squares of the negative ones. (Fourth powers might be better.) If the negative total is larger, the residue is inverted. Then, since only positive peaks will be used, the signal is half-wave rectified (negative values are set to zero). The rectified smoothed residue is shown as the third line of Figure 2.

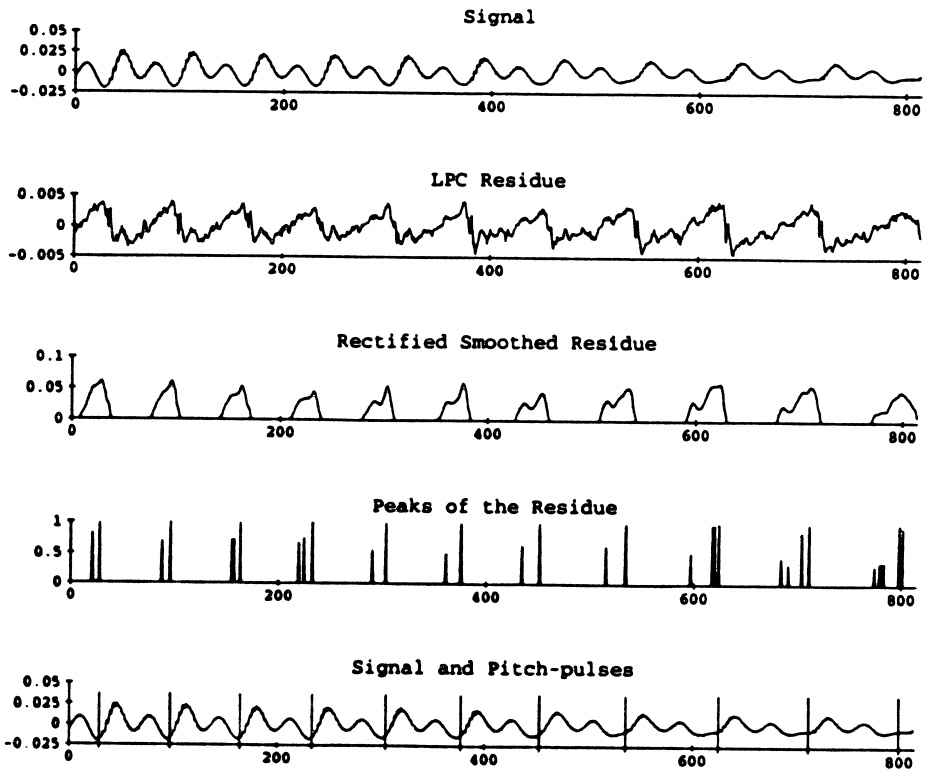


Figure 2: Steps in finding pitch epochs

4.2 Crude Pitch

Pitch is computed every 100 samples, using as input signal the low-passed LPC residue. The pitch finder is a correlation pitch detector with a dynamic-program pitch tracker, which returns at every time the "best" 11-long sequence of pitch periods ending at that time. Graphs of pitch tracks on all the material used in this study show that the pitch program made no egregious error on any token.

4.3 Finding Peaks

Next, all the peaks in the rectified residue are located and normalized. Normalization consists in expressing the height of a peak as the ratio of its height to that of the tallest peak within a span of the local pitch period. The fourth line in Figure 2 shows the normalized peaks.

4.4 Dividing into Epochs

Pitch and normalized peaks are passed to a dynamic-program "pitch-pulse finder", which accepts a sequence of candidate peaks (times and amplitudes), and finds the "best" subsequence, where goodness is a function of both consistency of the interval between chosen peaks, and their amplitudes. This subroutine, too, produces no egregious error on any token. The last line in Figure 2 shows the speech again, with pulses superimposed. There is one pulse per epoch.

4.5 Refining the Pitch Estimate

The last step in the algorithm is a more precise determination of pitch period. At each chosen pulse, as descibed above, there is a local (crude) pitch period P. At that pulse the signal is autocorrelated at every (integral) lag L in the range (P-P/2) to (P+P/2). More precisely, each pitch lag L is tested by forming the dot product of the L-long stretch to the left of the pulse with the L-long stretch to the right of the pulse, normalized by the power in the two L-long stretches. Figure 3 illustrates the process. Panel (a) panel shows the signal, centered at the local "pulse". Panel (b) is the autocorrelation function of this signal, computed at integral lags. (For a discussion of this technique see e.g. Hirose et al. 1992.)

Finally, cubic spline interpolation is done near the maximum of the autocorrelation function, creating values every 10 microseconds. Panel (c) is a blowup of the region inside the rectangle in panel (b), showing the interpolated points. The abscissa at which the maximum occurs is *defined* to be the pitch period.

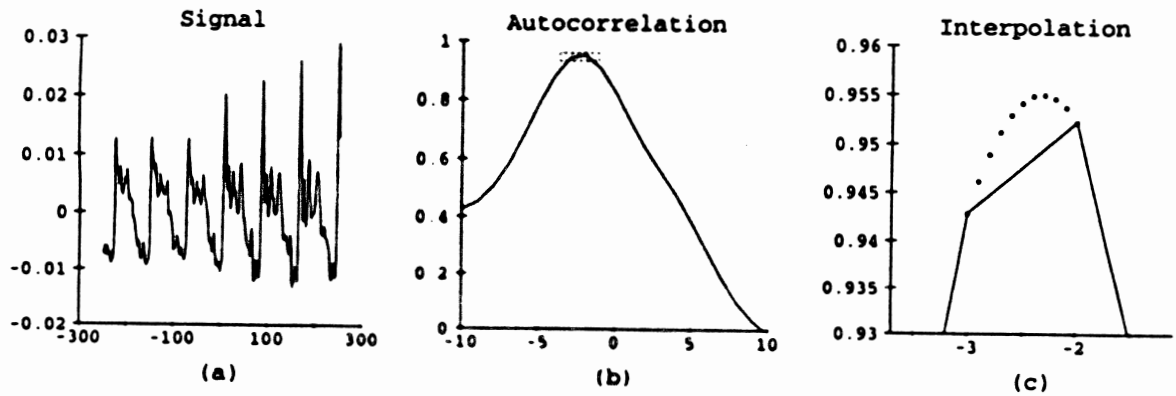


Figure 3: Refining the Pitch Estimate

5. Defining Jitter

Now that successive pitch periods for a token are known, we would like a number that indicates how much the sequence of pitch periods departs from a "smooth" sequence. There is little agreement among voice analysts on how to measure "roughness" in such sequences, even in sustained monotone vowels. (See e.g. Karnell et al 1991.) The simplest measure is just sums of (absolute) differences between adjacent pitch periods; this puts all its emphasis on adjacent epochs, and none on epochs two or more apart.

(Voice analysts are not alone having no obviously correct way to express roughness in a sequence of observations. Every expermental science that produces data of this kind has this problem, and many a statistician has blunted his spear (or obtained a Ph.D.) trying to find a satisfactory answer to this ill-posed question.)

Four definitions of jitter have been looked at to date; they give different numbers on any one token, but when used in a comparative way they all seem to tell about the same story.

5.1 Distance from average of adjacent values

A simple procedure, found in many papers on jitter, is based on absolute difference between the period now and the average of the previous period and the follower period. That is, if three successive pitch periods are $P(n-1), P(n)$, and $P(n+1)$, the local contribution to jitter is

$$J_1(n) = |P(n) - (P(n-1) + P(n+1))/2|$$

If epochs are short, then a given difference from expected is perhaps more significant than if they are long; one way to take this into account is to express the difference as a percentage of the local pitch period, and define jitter as the average of these percentages. Our first measure of jitter then, is

$$Jitter = (1/n) * \sum_n [100 * J_1(n)/P(n)]$$

5.2 Distance from line between adjacent points

Distance from average has a disturbing property. In places where pitch is changing fast, the contributions to total jitter tend to be larger those from regions of slow change. In Figure 4, the solid curve is pitch period, and the dotted one is the average of the two surrounding pitch periods. The vertical distance of the point at epoch 8 to its "expected" value is large, but it is really quite close to the line joining its two neighbors. (The vertical distance can be reduced by averaging in the current period as well as its two neighbors.)

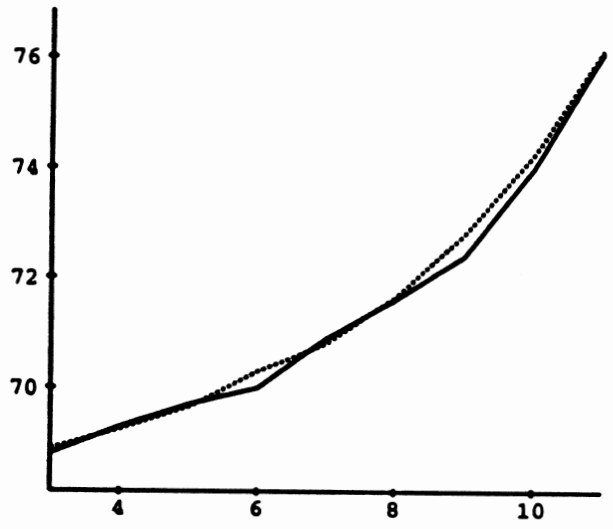


Figure 4: Pitch period vs. "expected" pitch period

To overcome this difficulty, one can take as the contribution of each point to the total jitter its *distance from the line joining its two neighbors.* If three consecutive periods are again $P(n-1), P(n)$, and $P(n+1)$ this distance is

$$J_2(n) = \frac{|(P(n+1) + P(n-1)) - 2 * P(n)|}{\sqrt{1 + (P(n+1) - P(n-1))^2}}$$

It is now perhaps inappropriate to normalize by the size of the local pitch period. Our second working definition is

$$Jitter = (1/n) * \sum_n J_2(n)$$

5.3 Deviation from a smooth curve

There are many algorithms for approximating a sequence of points by a "smooth curve". Use of any particular algorithm is based on a desire by the owner of a set of data to capture some feature of the data. In a recent paper on characterizing intonation patterns in spoken Mandarin, S. Chen and Y. Wang (Chen and Wang, 1990) describe an algorithm they consider appropriate to their problem, based on Legendre polynomials. On the sequences of pitch periods for the tokens in our study, except in a few instances their algorithm produces a curve that, to the eye, is indeed a "smooth" version of the plotted data.

Figure 5 shows three such curves (dotted line) superimposed on the data they purport to approximate (solid line). Panel (a) shows an unusually good fit, panel (b) an unusually bad fit, and panel (c) a typical fit.

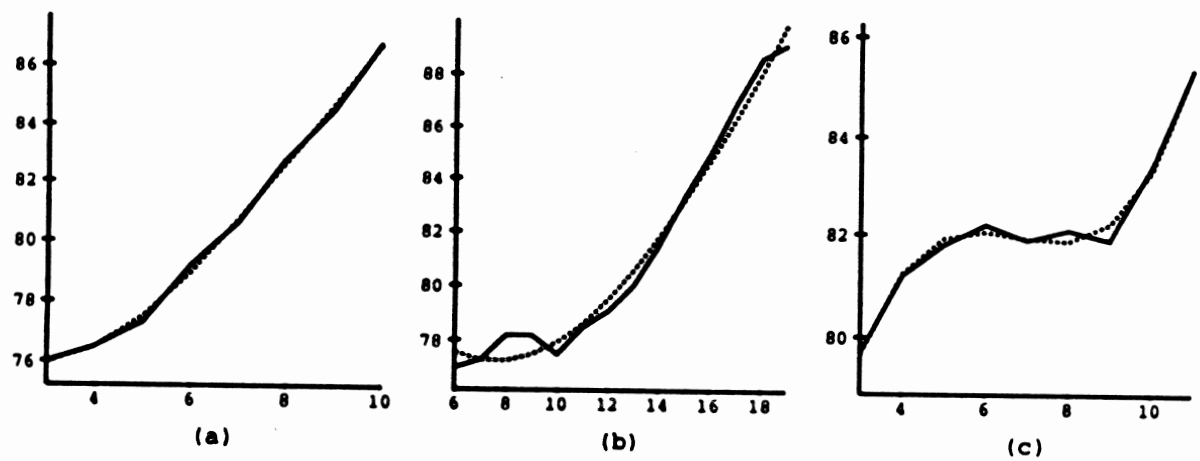


Figure 5: Approximating pitch periods with a smooth curve

If we call the ordinates of the approximating curve $\{C(n)\}$, and view $C(n)$ as the "expected" value of the nth period, then the distance analagous to that defined in Section 5.1 is

$$J_3(n) = |C(n) - P(n)|$$

Again by analogy with the procedure in 5.1, jitter is defined as

$$Jitter = (1/n) * \sum_n [100 * J_3(n)/P(n)]$$

5.4 Distance to a smooth curve

The fourth and final measure is again a response to the problem of having larger deviations where the pitch periods are changing rapidly. The program finds the distance from the plotted data point $P(n)$ to the *nearest* point on the curve $\{C(n)\}$, and jitter is the average of the absolute distances (not normalized).

6 Results

This is not a report on the results of the study on aging. What is of interest here is the consistency of a jitter measurement, that being our principal desideratum. Table 1 shows average jitter (of the four kinds) for each talker, at each age, saying each vowel.

			Jitter type			
Talker		Vowel	1	2	3	4
young	AH	AA	0.3754	0.1961	0.2592	0.1349
old	AH	AA	0.3906	0.2207	0.4337	0.2308
young	AH	IY	0.3598	0.1636	0.2151	0.0932
old	AH	IY	0.3905	0.2026	0.3244	0.1513
young	AH	UW	0.2768	0.1420	0.2065	0.0973
old	AH	UW	0.2237	0.1174	0.2467	0.1216
young	JM	AA	0.1937	0.0983	0.1696	0.0857
old	JM	AA	0.2835	0.1184	0.2564	0.1072
young	JM	IY	0.3426	0.1571	0.2556	0.1141
old	JM	IY	0.3711	0.1346	0.2404	0.0795
young	JM	UW	0.2092	0.0908	0.2053	0.0843
old	JM	UW	0.2529	0.0819	0.2675	0.0841
young	KS	AA	0.4692	0.2390	0.3302	0.1612
old	KS	AA	0.2848	0.1570	0.2974	0.1456
young	KS	IY	0.4511	0.1776	0.2863	0.1071
old	KS	IY	0.4446	0.1826	0.2443	0.0921
young	KS	UW	0.3335	0.1321	0.2141	0.0779
old	KS	UW	0.3226	0.1353	0.1719	0.0712

Table 2: Average of the four kinds of jitter, per talker, per age, per vowel

Even without statistical analysis, a few things are already clear:

Old KS shows less jitter than young KS;

Old AH shows more jitter than young AH;

AH and KS show more jitter than JM;

In many cases, amount of jitter is vowel-dependent.

7 Conclusions

The pitch-epoch finding algorithm is adequate. It is large and computationally intensive, but that is not a problem on today's computers. The author would be happy to share code (some Fortran, some C) for this or any other part of the computations described above.

For our purposes, there seems little to choose among the four measures of jitter described in this paper. Each is fairly consistent across a given condition, and they are fairly consistent with each other. We will continue investigating all four measures, and perhaps try others, on the rest of the collected data.

It would have been good to make direct digital recordings rather than tape recordings. Wow and flutter are probably not affecting average jitter (the law of large numbers is in our favor), but they surely affect the variance of the jitter. (Variance measurements on the small amount of data in this study are not encouraging.) This makes it fruitless to try to measure jitter on small samples, such as one phoneme in one context by one talker.

8 References

Chen, Sin-Horng and Wang, Yih-Ru (1990), Vector Quantization of Pitch Information in Mandarin Speech, *IEEE Trans. Comm. Vol. 38, No. 9*

Hirose, K., Fujisaki, H., and Seto, S. (1992), A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag, *Proc. ICASSP 1992 pp. I-149 to I-152*

House, A.S. and Stevens, K.N. (1993) Speech production: Thirty years after, *Jour. Acoust. Soc. Amer. Vol. 94, No. 3, Pt. 2 (Abstract)*

Karnell, M.P., Scherer, R.S., and Fischer, L.B. (1991), Comparison of Acoustic Voice Perturbation Measures Among Three Independent Laboratories, *Journal of Speech and Hearing Research, Vol. 34*

Titze, I.R. and Liang, H. (1993) Comparison of F_0 Extraction Methods for High-Precision Voice Perturbation Measurements, *Jour. Amer. Speech and Hearing, Vol. 36*

How we do it: Automated Target Matching and Data Selection Procedure in Voice Sample Acquisition: (Jack Jiang, David Hanson, Jie Chen, Northwestern University Medical School, Chicago, IL, 60611)

Introduction:

Inadequate, or contaminated samples are one of the common sources of error in clinical laboratory testing. For acoustic analyses of a vocal signal, it is important to obtain representative samples of the subjects natural range of frequency and intensity. Because the human voice can be voluntarily controlled within a wide range, it is essential to account for variables of frequency and intensity as well as other factors that may influence the sample.

Frequency and intensity are two important factors, which we know have some effect on phonatory physiology. The results of vocal acoustic analyses such as jitter and shimmer vary with the frequency and intensity of phonation. Pabon (1991) demonstrated that jitter and shimmer may vary significantly with different combinations of frequency and intensity. Therefore when obtaining samples of phonation, frequency and amplitude of the phonation should be defined in a fixed range (which may be considered a target area in the subjects phonatogram) in order to reduce the variance of measures. One way of obtaining samples of defined frequency and intensity is to ask the subject to phonate at specific frequency and intensity while providing some

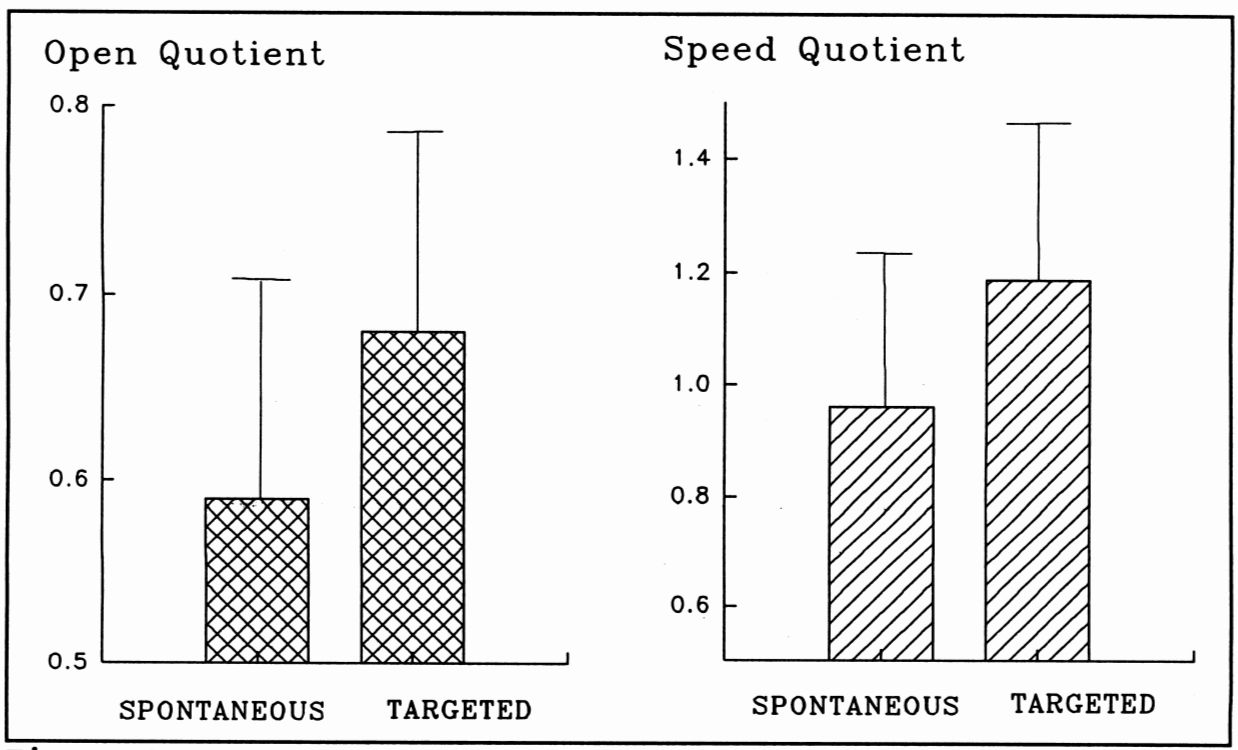


Figure 1

feedback. This has been called target matching (Hanson, et. al 1990). However, we cannot assume that frequency and intensity are the only, or even the most important variables in vibratory biomechanics during phonation.
In studies of speed quotient and open quotient of the vibratory cycle, measures of phonation that was matched to specific frequency / intensity targets were compared with spontaneous phonation of comparable frequency and intensity. It was found that the effect of matching specific frequency intensity targets influenced change in Speed Quotient(SQ) and Open Quotient(OQ) to a greater degree than these measures changed across the range of the subjects frequency and intensity. In other words the variable of target matched (versus spontaneous production) was greater than variation from the lowest to highest pitch and loudness. As seen in
figure 1 (Hanson 1990), the open quotient and speed quotient of targeted phonation was significantly greater than for spontaneous phonation at the same frequency and intensities.

Perturbations also appear to be effected by the effects of target matching. As the figure 2 shows, jitter and shimmer were also greater for targeted phonation than for spontaneous phonation of the same frequency and intensity. Therefore, the task of matching specific frequency / intensity targets for sampling phonation may introduce possible error into the samples if they are to be compared to spontaneous phonation. In this study, we report how in our voice laboratory we have attempted to reduce

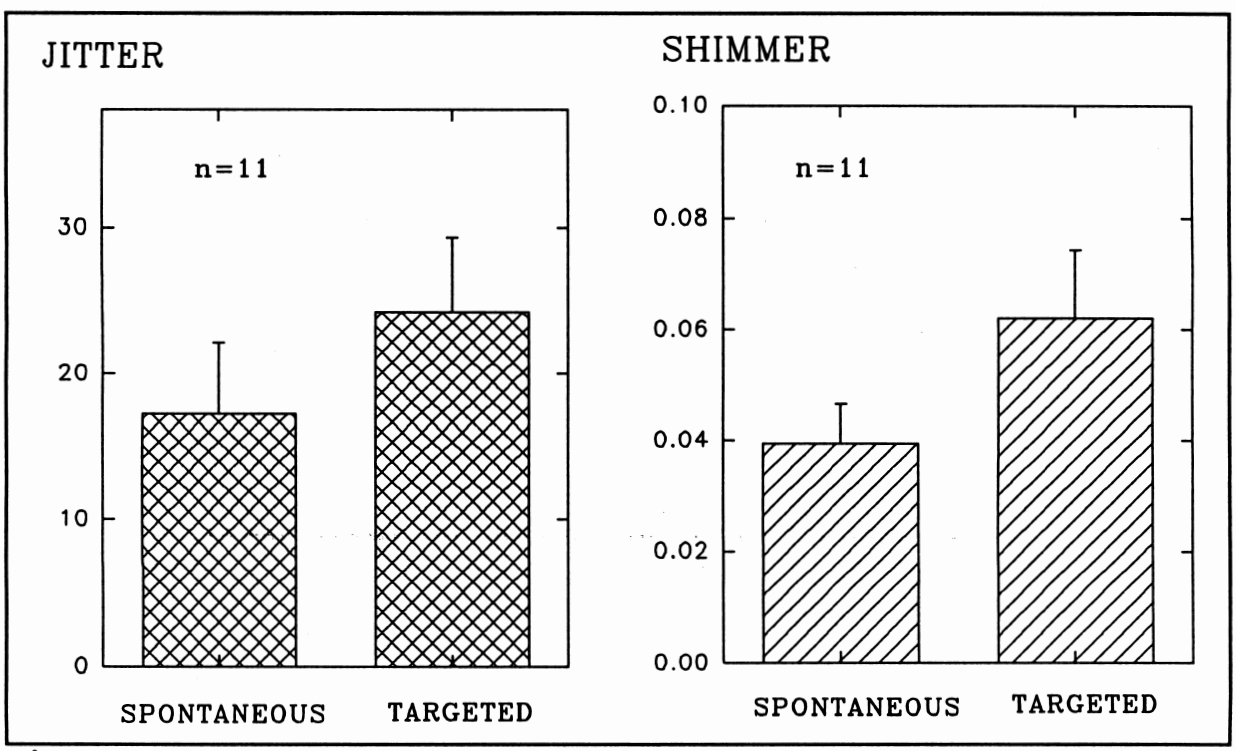


Figure 2

this problem in sampling phonation by a target matching method.

Methods:

Target Matching Technology

Figure 3 illustrates the data acquisition system which directly related with the sample selection in this study. An AKG Boom Set C-410 condenser microphone was connected to a Symetrix SX 202 Pre-amplifier for recording the acoustic signal. The output of the pre-amp was connected to a root-mean-square(RMS) to voltage converter. The output of the convertor is proportional to the logarithm of the RMS of the acoustical signal. Both output of the RMS and the output of the preamplifier were sent to A/D board for digitizing. A Quest model 2800 Sound Pressure Level Meter was used for intensity calibration.

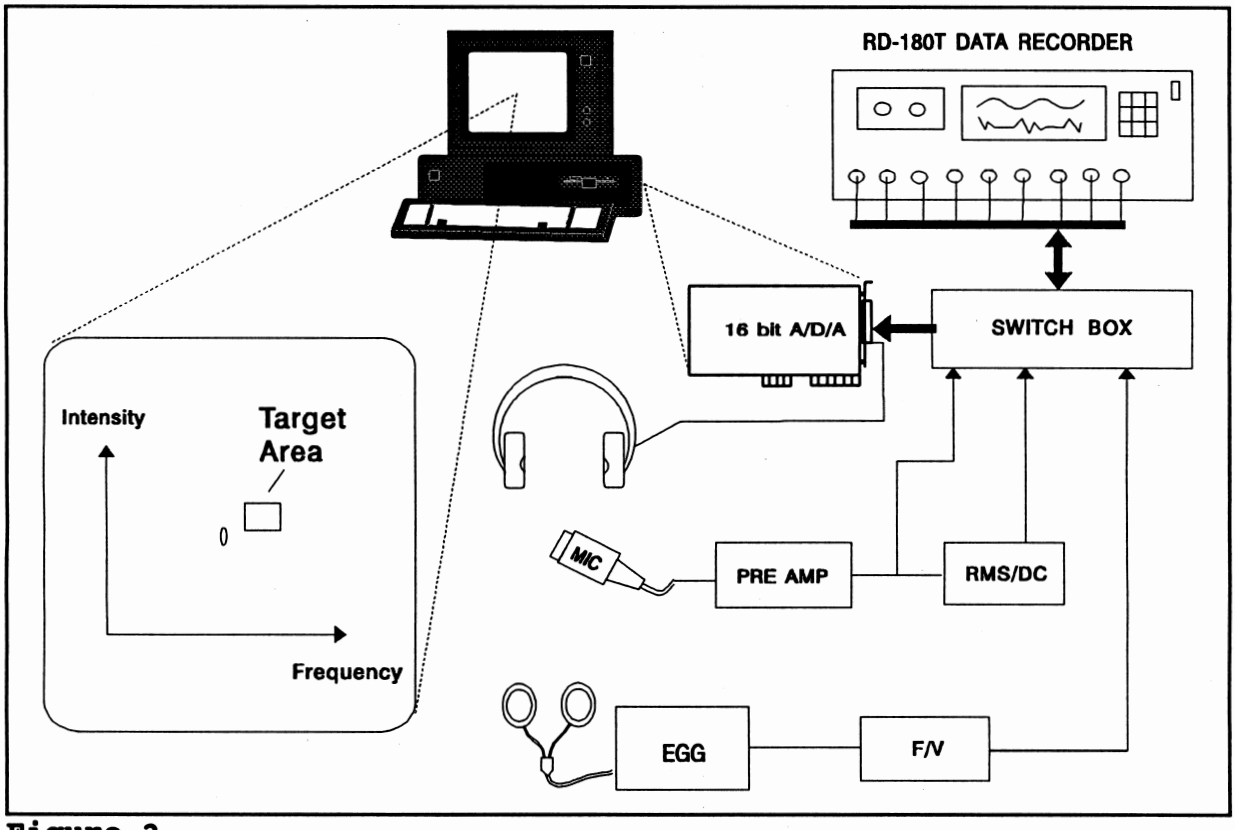


Figure 3

Frequency was obtained from the electroglottogram. A Glottal Enterprises (Rothenberg's) SC-1B Single Channel EGG unit was used. The electrodes were placed on the either side of the thyroid alae. A frequency to voltage convertor was used for providing a voltage proportional to the frequency of the vocal fold vibration, which was equal to the fundamental frequency of

voice.

A 33MHz 486 personal computer with two National Instruments AT-MIO-16F (12 bits 16 channel) A/D boards with aliasing filters was used for digitizing data. Labwindows 2.1 (National Instruments) was used for developing the data selection and analysis software. The sampling frequency, and the gain of each channel, was software selectable from a window environment. The data and the resulting files were saved in a pre-organized form so that data searching could be automated.

The data files, synthesized tones, or a pre-digitized voice instruction were played back to an earphone or the field speakers through two 12 bits D/A channels of the same board and power amplifier.

Before data acquisition, the examiner chose among the following parameters: a. Desirable range in intensity and frequency; b. The length of the segments needed; c. Sampling rate

In our custom made software, data acquisition has two modes that are based on the triggering strategy. These are a manual mode and an automatic target matching mode. In the automatic target matching mode, there are two steps for the data acquisition:

The first step was to determine the most comfortable intensity and frequency range of the subject. The subjects voice was recorded during casual conversation to obtain a spontaneous sample of speech for 20 seconds. The average frequency and average amplitude were then calculated from digitized frequency and amplitude contour data.

The second step was to select the data automatically based on frequency and amplitude of the phonation. The most comfortable intensity and frequency of the subject was used to determine the center of the target. Target area was determined by adding, and subtracting a pre-set range to the most comfortable intensity and frequency.

For target matched phonation, the subject was instructed to give a long and steady phonation. A synthesized target tone was played through a pair of speakers to the subject at the particular target. In addition to the target tone, a frequency contour and amplitude contour were digitized and displayed as x-y plots on the computer screen in real time for visual feedback, as shown in figure 4.

In order to capture spontaneous phonation the subject was encouraged to produce phonation of different frequencies at different intensities. During data acquisition, the digitized data (3 or 4 channels in total) was saved as a temporary file on computer hard drive. If frequency and amplitude contours were both in the pre-selected range in which we desired to obtain a sample, and lasted continuously for a pre-selected duration, such

as 2 seconds (see figure 5), then the data of last steady section of phonation in temporary file was saved as a file with a given name. The rest of the data in the temporary file was discarded. If the frequency and the amplitude was not steady for 2 seconds, the program reset itself every 120 seconds by rejecting all the data in the temporary file and starting a new one.

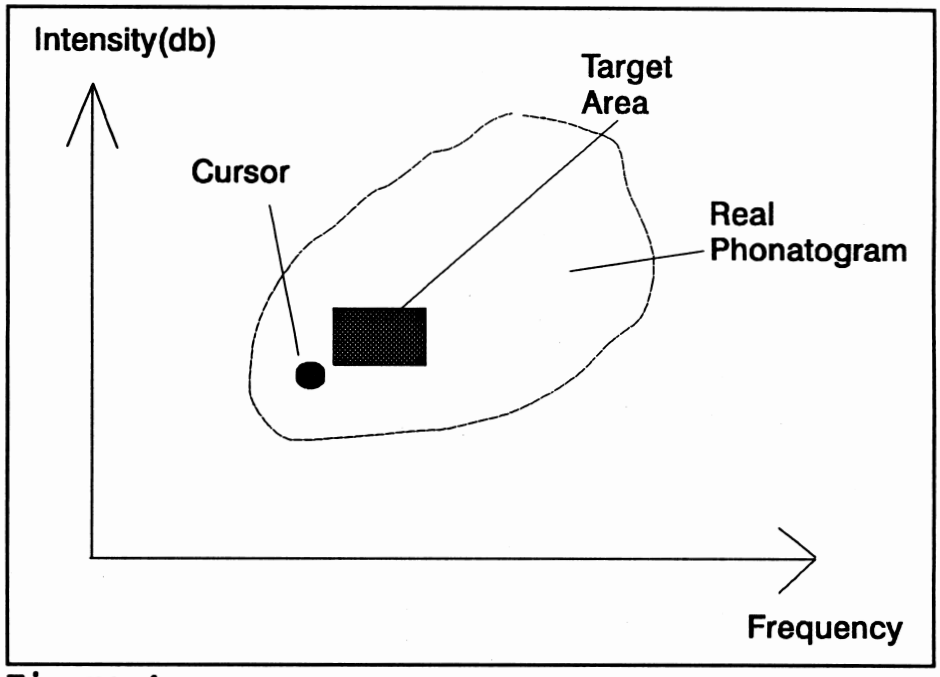


Figure 4

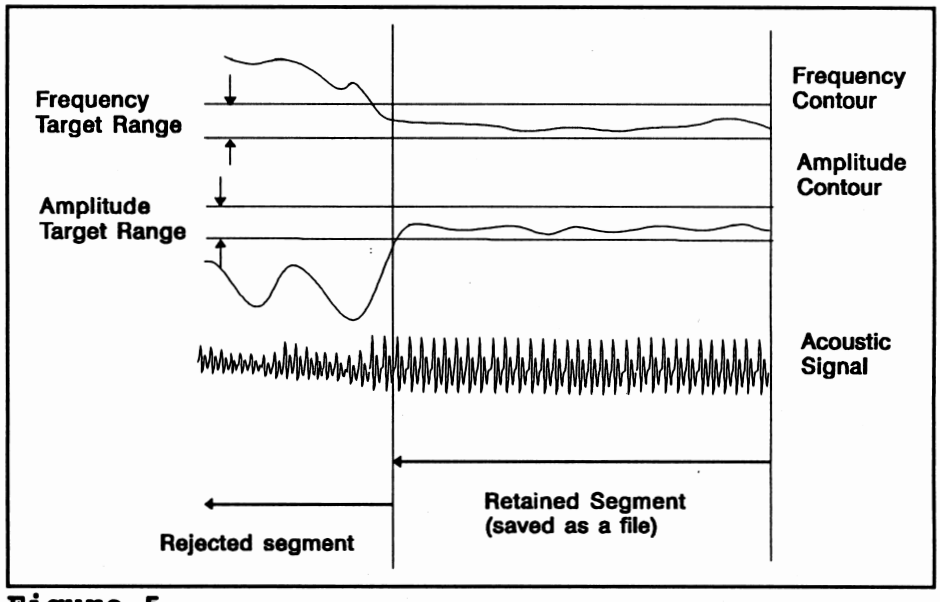


Figure 5

Backward Data Acquisition

One convenient feature of our system was its "backward" data acquisition. When the investigator wanted to acquire a section of patient voice data with specific characteristics, it was difficult, or tedious to do in a conventional system. Because the investigator did not always know ahead of time when the right combination of frequency and intensity would be achieved it was necessary to have the acquisition time long enough to wait for the sample that was desired. After the acquisition, it was necessary to search, what was sometime a huge file, for the desired sample, another tedious task.

We automated this procedure by using the "backward" acquisition

mode. In this mode, the investigator first specifies the length of the section he/she wants to acquire, and starts the acquisition in the background. From this moment the data is stored in a temporary file. As soon as the physician recognized the desired piece of data he/she wants, he/she can hit the RETURN key to stop the acquisition. The program will then find the end of the temporary file, and will backward trace to the length previously specified and will copy this section of data to the final data file(figure 6).

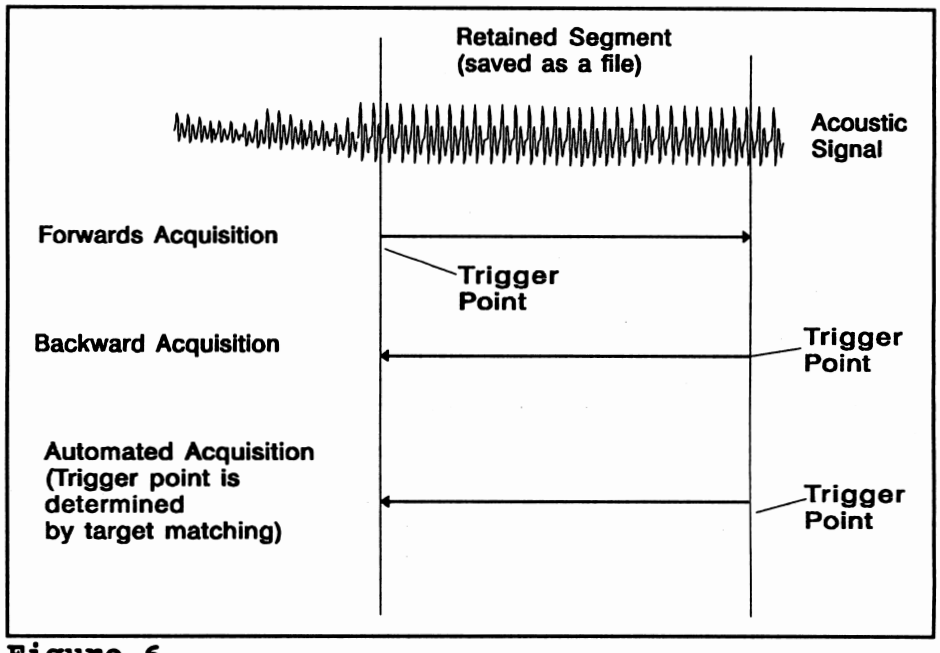


Figure 6

The Double Buffer Technology

To acquire large blocks of data at high speed, the double buffer acquisition technique was used. Usually, the hard disk of the computer was much larger than the RAM (Random Access Memory). But the disk access speed was too slow for high speed data acquisition. RAM was much faster for acquisition but of limited size for storage. The double buffer technique let us combine the best of the two. The maximum sample rate was almost as high as a single memory buffer mode, and its maximum size was only limited by the hard disk space. The method involves two memory buffers, allocated as a data buffer and a transfer buffer. The data buffer was used to receive data from the A/D converter at a fixed sample rate. It was configured as a cyclic buffer, i.e. if the buffer was full, the new data overwrote the previous data from the beginning of the buffer. The transfer buffer was half the size of the data buffer. When more than half of the data in the data buffer were new (i.e. have not been copied to the transfer buffer), the program signaled a flag that it was ready to copy the new half data buffer to the transfer buffer. The data in the transfer buffer were then written to a file in hard disk. The copy of half of the data buffer to transfer buffer and the write of transfer buffer to disk file do not interrupt the data acquisition carried on in the data buffer. This block transfer technique utilizes both the high access speed of RAM and the large memory of hard disk space to achieve the high speed acquisition of the large block of data. One caution in using the

"double buffer" technique was the possibility that desired data could be overwritten. If the program does not check the "data ready" flag of the data buffer often enough, the incoming data might overwrite the data that have not been copied to the transfer buffer and these data were permanently lost. One should test this for the data acquisition board and computer system to be used. The system we use (National Instruments MIO-16F Board with Gateway 486, 33MHz with 15 ms HD) can perform this reliably up to 100 K sample rate.

Discussion:

It has been reported that frequency and amplitude data can be obtained by a TMS 32010(Pabon 1991) DSP chip or a more user friendly TMS 320C30 digital signal processing board(Titze 1993). In our setup, the frequency and amplitude were both converted as analog voltages by a few IC chips. The rational for this was to avoid the technical difficulty of using a DSP chip and also to reduce the cost. The cost for develop both RMS/DC, and F/V converters was less $100.

The RMS/DC(AD536) convertor has a 60 db dynamic range. The average time constant was determined by the value of capacitor. The response time was 0.1 second in out design. The error was <1% at the signal crest factor of 7. For the F/V convertor, the maximal input frequency was 1500 Hz. The response time of the F/V converter was set to 100 msec. The linearity error was than 0.3%.

Our preliminary experience was obtained with a target size that was 10Hz x 5 db around the most comfortable frequency for a normal subject. It is obvious that the narrower the range, and the rougher the voice quality, the more difficult it will be obtain a sample of the desired range. The target range for pathological voices will probably need to be larger and will be determined experimentally in the future. Because of high frequency components (high crest value) of the acoustic voice signal, it is not accurate to use a F/V convertor to determinate F0 based on the acoustic signal. The fundamental frequency obtained from the EGG signal is more accurate.

The effort of target matching effects several measurable characteristics of phonation. The mechanism for this is not known but is probably related to an increase in effort or strain. While spontaneous phonation is a comfortable relaxed practiced procedure, matching voice to a specific frequency and intensity target is for most untrained subjects an unfamiliar difficult procedure. Muscle tension and mild tremor appear to be greater with such effort. This would account for perturbations that were higher in phonation produced while matching targets than they were in spontaneous phonation of similar frequency and intensity. This hypothesis seems to be supported by measures of increased speed quotient, open quotient, jitter and shimmer seen in target matching phonation samples.

To be able to obtain samples of specific desired frequency and amplitude and, at the same time to reduce the effort of target matching was a challenge for our studies of the vibratory physiology of the vocal folds during phonation. Our preliminary experience indicates that auditory feedback of the tone that we want the subject to match, is easier for most subjects that asking them to match a target from real time visual feedback. Encouraging the patient to produce a variety of samples, (" a little higher... a little louder") and having the sampling program keep track of when the sample is in the desired range for capture, we expect to see less perturbation in acoustic recordings and physiologic measures. We are currently studying these phenomena.

Reference:

Pabon JPH (1991), Objective acoustic voice-quality parameters in the computer phonetogram, Journal of Voice, Vol.5, No.3 pp 203-216

Hanson DG, Gerratt BR, Berke GS (1990), Frequency intensity and target matching effects on photoglottographic measures of open quotient and speed quotient, Journal of Speech and Hearing Research, Volume 33, 45-50.

Measures of Vocal Function During Changes in Vocal Effort Level

Daniel Zaoming Huang

Fred D. Minifie

Department of Speech and Hearing Sciences, JG-15

University of Washington, Seattle, WA 98195, USA

Hideki Kasuya

Department of Electronic Engineering

Utsunomiya University, Utsunomiya 321, Japan

Sarah Xiao Lin

Tiger Electronics, Inc.

P.O.Box 85126, Seattle, WA 98145, USA

Summary

The purpose of this paper is to present the results of a controlled study of the day-to-day variabilities of three acoustic parameters (jitter, shimmer, and normalized noise energy), and two electroglottographic (EGG) parameters (contact quotient and contact quotient perturbation) for vowels produced at three vocal efforts (soft, normal, loud). Data were obtained using a sophisticated bilinear interpolation pitch detection method. A repeated measures design required

subjects to produce the vowels /ae/ and /a/ five times a day over three days at each vocal effort level. The jitter, shimmer, and normalized noise energy (NNE) values from acoustic measures and Contact Quotient (CQ) and Contact Quotient Perturbation (CQP) values varied significantly among the three vocal effort levels. The clinical implication of this finding is that vocal effort must be controlled in order to obtain consistent clinical measures. Furthermore, day-to-day variability must be taken into account if representative measures are to be obtained for clinical use.

Key Words

Jitter, shimmer, normalized noise energy, contact quotient, contact quotient perturbation, vocal effort, pitch detection.

Introduction

Scientists have long known that clinical use of acoustic and electroglottographic (EGG) measures provides a convenient and non-invasive way to evaluate laryngeal function (Davis, 1976; Aronson, 1980; Huang and Hu 1988). The three acoustic measures that have received the most attention in the literature as indicators of vocal function are cycle-to-cycle variations in fundamental period (jitter), cycle-to-cycle variations in peak-to-peak amplitude (shimmer) and normalized noise energy (NNE) (Hirano, Matsushita, and Hiki, 1976; Kasuya, Ogawa, and Kikuchi, 1986). There are four EGG measures that are also provide useful information about normal and pathological vocal function. The four EGG measures are contact quotient (CQ), contact quotient perturbation (CQP), cycle-to-cycle fundamental period variations from EGG signal (EGG-jitter), cycle-to-cycle peak-to-peak variations from EGG signal (EGG-shimmer) (Baken, 1987; Huang 1988; Huang, Minifie, & Lin 1992). The usefulness of such measures as indicators of vocal function is dependent upon the irreliability and the sensitivity of the measures to changes in vocalizations. Previous studies have looked at intrasubject variability of vocal jitter in voice signals from day to day (Linville, 1988; Haggins and Saxman, 1989), the relationship of

vocal jitter to voice intensity levels (Titze, Horii & Scherer, 1987), vocal jitter changes with the aging voice (Brown, Morris, and Michael, 1989), and differences in vocal jitter from vowel to vowel (Orlikoff and Huang 1991). Similar studies need to be done to indicate the relative stability of each of the acoustic measures and EGG measures used to evaluate vocal function.

Accurate characterization of acoustic and EGG measures is essential not only in the evaluation of vocal pathologies, but also in the accurate modeling of the voice source for the speech synthesis (Fant, 1980). Two of the major questions about acoustic measures and EGG measures remain unresolved: 1) how do these measures change with changes in vocal effort level, and 2) what is the day-to-day variability in these measures ? One way to address these questions is to investigate intrasubject patterns of variation in a group of normal subjects. A better understanding of the variability of voice perturbation of normal speakers at different vocal efforts over time is needed, therefore, before the use of acoustic measures and EGG measures can be used appropriately as clinical measures for voice assessments.

There are three purposes for this study: 1) to introduce a sophisticated bilinear interpolation pitch detection method, 2) to use the new pitch period detection method to evaluate the stability of acoustic and EGG measures of vocal function during changes in vocal effort level, and 3) to evaluate the stability of such measures from day to day.

The Method of Analysis

A. Algorithm and Terminology

It is crucial to have an accurate cycle-to-cycle pitch period detection method, because nearly all aspects of acoustic and EGG measures are based on the accuracy of pitch detection. For example, since perturbation measures rely on the accurate identification of each pitch period, measures of jitter and shimmer are dependent on the precision of pitch period detection.

Similarly, the detection of glottal noise requires an accurate pitch period marker in order to match adjacent waveshape cycles.

Various F0 extraction methods have been summarized by Hess (1983). They can be classified in two major categories: 1) event-detection methods, such as the peak-picking and zero-crossing methods; and 2) short-time averaging methods, such as autocorrelation, minimal distances, amplitude magnitude difference function, cepstral analysis, and harmonic compression. Milenkovic (1987) found that greater reliability and accuracy could be obtained by matching the entire waveshape across adjacent cycles rather than by identifying isolated events, like zero-crossing and peak-picking. Kasuya (1986, 1989) developed a rather accurate pitch detection method based on a cycle-to-cycle amplitude magnitude difference function (AMDF). This pitch detection method compares the sampled data points for an entire waveform with those from adjacent cycles.

Since the measures of jitter, shimmer, NNE, CQ, and CQP are based on the cycle-to-cycle similarity of the wave form, accurate determination of the pitch period is of crucial importance. This paper presents a new method for determining pitch periods, from which jitter, shimmer, NNE, CQ and CQP are measured. The method incorporates a bilinear interpolation procedure into the average magnitude difference function (AMDF) to evaluate the cycle-to-cycle waveform similarity in sustained vowel utterances. This method was selected based on experiments with synthetic speech showing that the method performs better than several other methods taken for comparison. The method of bilinear interpolation of sample points on the Amplitude Magnitude Difference Function (AMDF) is shown in figure 1. The pitch markers ($q_1,q_2,q_3,\ldots,q_{n-1},q_n$ or $c_1,c_2,c_3,\ldots,c_{n-1},c_n$) shown at the top of figure, are estimated using an automatic method based on zero crossings of the vowel wave form. The method then locates the pitch boundary on the basis of the AMDF to indicate the beginning of each pitch period. In this case, six points, shown at the bottom of figure, around a primary dip in the AMDF are separated into two groups; one group includes a minimum AMDF point, while the other includes the second minimum point. Two lines

are obtained from the two groups on the basis of the least mean square criterion. The point where two lines cross is regarded as the real pitch boundary, the beginning of a real pitch period $(P_1,P_2,P_3,...,P_{n-1},P_n)$.

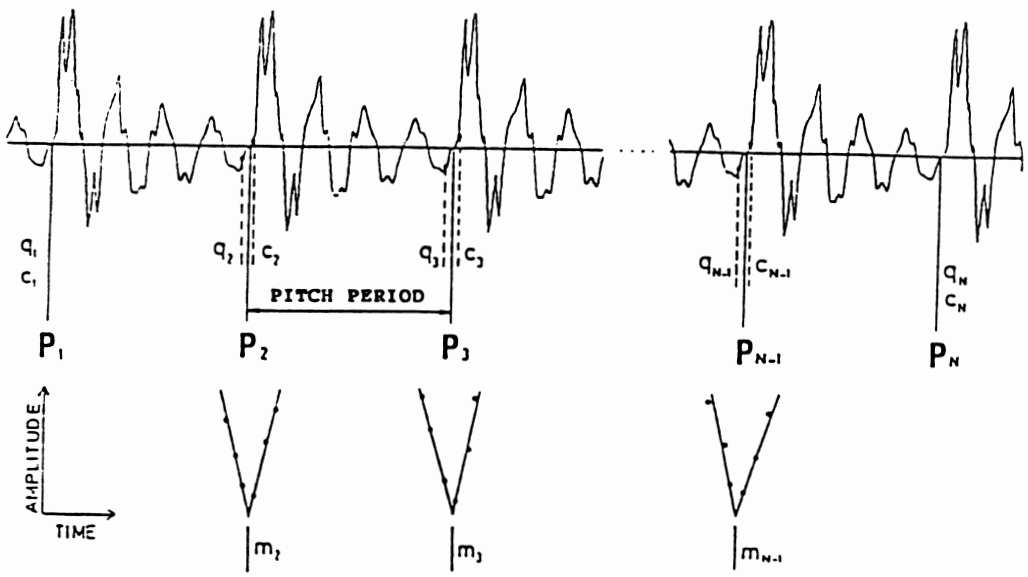


Figure 1. Schematic illustration of the pitch detection method with bilinear interpolation of the average magnitude difference function (AMDF).

For a sequence p(n), n=1,2,...,N, the perturbation quotient PQ (%) is defined as

$$PQ = \frac{100}{N-k+1} \sum_{n=1}^{N-k+1} \left| 1 - \frac{k * p(n+m)}{\sum_{j=1}^{k} x(n+j-1)} \right| \qquad (1)$$

where k is the length of moving average (an odd integer greater than one) and m=(k+1)/2. In our system k=5 and m=3. If p(n) is the pitch period of acoustic signal, then PQ is the pitch period perturbation quotient (jitter), and if p(n) is the peak-to-peak amplitude of acoustic signal, then PQ

is peak-to-peak amplitude perturbation quotient (shimmer). If p(n) is the contact quotient sequence of the EGG signal, then PQ is the contact quotient perturbation quotient (CQP).

These jitter and shimmer values are measured from the pitch period and peak-to-peak amplitudes, respectively. After computation of the perturbation measures, the pitch period of each glottal vibration included in the count was displayed. More than 50 cycles were used for each perturbation analysis as supported by Titze (1987). Only segments that had pitch period fluctuations within 10% in either a positive or negative direction of the mean pitch period were analyzed. This criterion was used so that only very steady wave form segments would be analyzed for all subjects, thus minimizing variability due to selection of cycles for analysis. If no segment consisting of at least 50 cycles could be found to fit this criterion, no perturbation values were computed for that vowel production. Only a few of the normal vowel prolongations were rejected by following this criterion. This criterion is more problematic in analyses of pathological voices because many abnormal voice have relatively few stable segments with pitch period fluctuations within 10% of the mean pitch period. Such extremely variant voices cannot be analyzed. Accuracy of the new pitch detection method will be discussed with synthesized steady vowels later.

With respect to the EGG-jitter and EGG-shimmer, Haji (1986) found that the EGG-jitter was nearly equivalent to the jitter and shimmer obtained from acoustic signal, so that the EGG-jitter and EGG-shimmer data are not reported in this paper. The CQ measure from EGG signal provides unique information about vocal fold behavior that is, for the most part, invisible to other available techniques (Baken 1987; Orlikoff and Baken 1990). The CQP measure provides precise information about the rate, symmetry and regularity of the vocal fold contact phase during vocal fold vibration (Huang, Minifie and Lin 1992, 1993). It is for these reasons that the CQ, CQP measures were chosen in our study. Rothenberg (1988) has suggested the use of variable baseline crossings with interpolation of a criterion level to demarcate the EGG contact phase. In

the present study, a baseline of 25% of the peak-to-peak EGG amplitude of each wave is associated with the EGG minimal contact phase and is selected for measuring the CQ and CQP.

The method of noise energy measurement used in this experiment provides more insight into perturbation measurement. The relative magnitude of noise included in the voice signal is evaluated using an acoustic measurement NNE (normalized noise energy) described by Kasuya (1986). We have chosen to use the normalized noise energy measure because it can differentiate among normal and pathological voices more sensitively than does the harmonic-to-noise ratio (Kasuya 1993; Hirano 1989). An adaptive comb filtering method is used in NNE for estimating vocal noise in normal and pathological voices. (This procedure was initially investigated for the enhancement of degraded speech due to additive white noise.) The NNE (dB) is given by the equation:

$$NNE = 10\log\frac{\sum_n w(n)^2}{\sum_n x(n)^2} + BL \qquad (2)$$

where w(n) and x(n) are respectively an estimated vocal turbulent noise component and an original voice waveform, and BL is a constant for compensating for the amount of noise energy removed by the comb filter.

B. Test Signals

The accuracy of the pitch detection method was tested using periodic synthesized signals without additive noise, which were produced by the following equation:

$$y(n) = \sum_{k=1}^{M} A(k)\sin\left(\frac{2nk}{T} + \Phi(k)\right) \qquad (3)$$

where A(k) is the amplitude of k-th harmonic component which simulates a vowel /ae/ as in "bat", $\Phi(k)$ is the phase of k-th harmonic component, M is the number of harmonics, and T is the normalized pitch period (points). In our system, $\Phi(k)=0$ and M = 23. The T is defined by the following equation:

$$T = F_s \times P \qquad (4)$$

where F_S is the sampling frequency, and P is the pitch period. For simulating a child voice, T is allowed to vary from 133 to 134 points with a step 0.2, which corresponds to a change from 3.325 to 3.35 ms. For simulating a female voice, T is allowed to vary from 174 to 175 points, which corresponds a change from 4.35 to 4.375 ms. Similarly, for simulating a male voice, T is allowed to vary from 333 to 334 points, corresponding to a change from 8.325 to 8.35 ms.

C. *Accuracy of the pitch detection methods*

Results showing the accuracy of the pitch detection method, with and without interpolation, are provided in Table 1. Here, two interpolation methods on the AMDF were employed in order to determine which method provides the more precise pitch period extraction. The two methods are: parabolic interpolation, and interpolation with bilinear approximation. The measures obtained using these interpolation methods were compared to measures derived when no interpolation was employed.

The results obtained from these test signals, which include a constant pitch periods from integer multiples and non-integer multiples, allow us to draw the following conclusions about pitch period detection. First, the standard deviations of data obtained via the parabolic and bilinear interpolation methods were always smaller than the standard deviation when no interpolation was used. The second observation is that the bias of the interpolation methods was generally smaller than the bias obtained with the "no interpolation" method. Third, the bias of bilinear interpolation method was always smaller than the bias obtained with the parabolic method. Thus, it appears clear from Table 1 that the bilinear method is superior to the parabolic

method. Also, Table 1 indicates that both of the interpolation methods are better than the no interpolation method.

TABLE 1. Comparison of the accuracy of pitch extraction, with or without interpolation. Bias is the difference of an average of measured pitch periods from the actual value, and SD is the standard deviation of measured pitch period values.

Pitch Period (point)	Pitch Period Error (point)					
	No Int.		Parabolic Int.		Bilinear Int.	
	bias	SD	bias	SD	bias	SD
133.0	.032	.183	.001	.001	.000	.000
133.2	.007	.402	.077	.078	.004	.006
133.4	.013	.495	.069	.070	.001	.004
133.6	.013	.495	.065	.066	.005	.006
133.8	.007	.402	.073	.075	.009	.010
134.0	.032	.183	.001	.001	.000	.000
174.0	.032	.183	.001	.001	.000	.000
174.2	.020	.402	.072	.073	.012	.011
174.4	.013	.495	.064	.065	.010	.010
174.6	.013	.495	.070	.072	.007	.007
174.8	.007	.402	.078	.080	.004	.005
175.0	.032	.183	.003	.003	.000	.000
333.0	.032	.183	.001	.001	.000	.000
333.2	.027	.385	.076	.077	.001	.003
333.4	.021	.494	.067	.069	.001	.002
333.6	.014	.501	.066	.067	.002	.002
333.8	.007	.412	.074	.076	.002	.005
334.0	.035	.189	.001	.001	.000	.000

In order to estimate the sensitivity of the bilinear interpolation method of pitch measurement, white-noise signals were scaled appropriately and then added point-for-point with the above periodic synthesized signals y(n). As the signal-noise ratio (SNR) decreased, reflected by increasing the amount of white noise added point-for-point on synthesized signals, jitter and the value of normalized RMS error of the pitch period clearly increase, as shown in figure 2. also, it is clear that variations in jitter imposed by noise are relatively small when SNRs are more than

45 dB. These data provide support for the use of a bilinear interpolation of the AMDF as a pitch detection algorithm. It appears to be an accurate method for pitch period measurement.

Therefore, in this paper, pitch detection in both acoustic & EGG signals was obtained by incorporating bilinear interpolation of sample data points on average magnitude difference function. Perturbation measures were obtained using a 5-point moving average procedure.

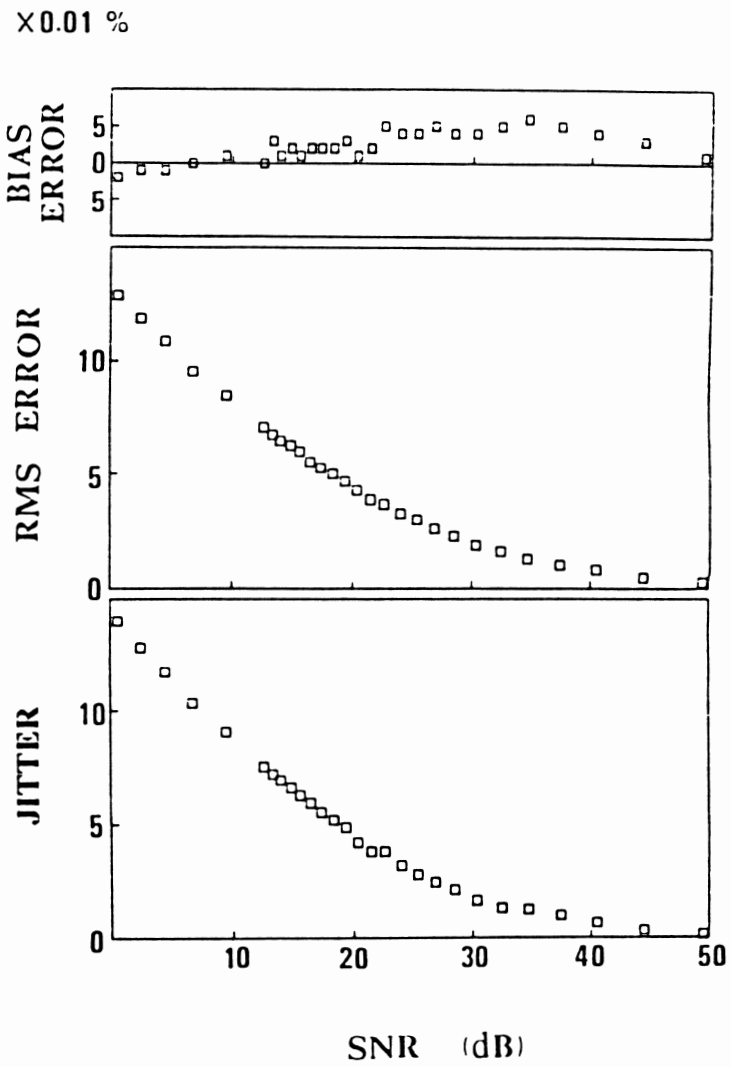


Figure 2. Jitter, normalized RMS and bias errors as a function of signal-to-noise ratio (SNR) of synthesized signals with white noise.

Voice Perturbation Measurements

The next investigation was to examine the influence of three vocal efforts on the measures of jitter, shimmer, NNE, CQ, and CQP over time. Three male subjects pronounced the sustained vowels /ae/ and /a/ at three vocal efforts (soft, normal and loud) five times a day on three different days. Jitter, shimmer, NNE, CQ and CQP were measured from each vowel sample.

While these acoustic and EGG measures may be easy to obtain, useful in analysis of the voice disorders and helpful in measuring progress during therapy, it is important to understand how these measures change during different vocalization conditions. Hence, our interest is how these measures change during voice production at different vocal effort levels.

A. Subjects

Subjects were three normal male adult subjects with no history of voice disorders, or present complaint of voice disorders. All subjects were in good health on each day of testing with no history of audiological, neurological or chronic respiratory disease.

B. Stimuli

We manipulated vocal effort level by having subjects produce the vowels /ae/ and /a/ at "soft", "normal", and "loud" vocal levels. Using a repeated measures design, each of three adult male normal talkers produced five replications of each vowel, at each vocal effort level, on each of three different days. These utterances were produced under two different conditions: 1) spontaneous vowel productions, and 2) imitative vowel production.

Spontaneous vowel production: Each subject was first directed to sustain the vowel /ae/ as in "bat" five times with each utterance lasting for more than 3 seconds at normal effort. Then, the subject was asked to repeat the sustained vowel five times again with each production lasting more than 3 seconds, at a soft vocal effort. Similarly, five replications of vowel were obtained at loud vocal effort. The same procedure was used to obtain tokens of the vowel /a/ at each of the three vocal effort levels.

Imitative vowel production: Each subject was required to sustain each vowel (/ae/ and /a/) in imitation of synthetically generated vowel tokens at each of the intensity levels: loud=75 dB SPL, normal=72 dB SPL, and soft=67 dB SPL.

Each subject for this experiment was seated in a sound-proof room (IAC 1200) and comfortably positioned in a head rest so that a condenser microphone (SONY ECM22-P) was positioned at a constant microphone-to-mouth distance of 10 cm. A throat-contact microphone was placed over the thyroid lamina to obtain electroglottographic signals. During the recording of both acoustic and EGG signals into a computer, no attempt was made to control fundamental frequency at any of the vocal efforts.

Each vocal token produced by the subjects in this experiment was digitized at a sampling frequency of 22050 Hz per channel with an accuracy of 16 bits/sample, and analyzed by using the software: Voice Evaluation and Therapy (VET 2.00) from Tiger Electronics (Huang, Minifie & Lin 1992). Only the middle portions of the vowel at each vocal effort were used for analysis.

C. Results

The results of this experiment can be seen in the following series of figures. Figure 3 shows the results for the vowel /ae/ produced in a natural, spontaneous manner. Each of the bar graphs shows the means and standard deviations of the data obtained for each vocal effort level condition: soft, normal, and loud. For example, it can be seen in Figure 3(a) that jitter decreases with increasing vocal effort level. Similarly, shimmer reduces with increasing vocal effort level (Figure 3(b)). Please note that we have used the acronym NNE (Kasuya 1986) to represent normalized noise energy (or what we have referred to above as glottal noise energy). Obviously this graph has to be interpreted in light of the fact that noise energy is measured in relation to the amplitude of the harmonic energy in the vowel. Therefore, the minus values indicate how many decibels below the signal energy is the level of the noise energy (e.g., a smaller minus value indicates a larger amount of noise than does a larger minus value. Figure 3(c) shows that as vocal effort level is increased, that the relative amount of noise in the vocalization decreases.

If we look at the Contact Quotient graph (Figure 3(d)), it can be observed that at loud levels of phonation, the vocal folds are closed for a considerably longer percentage of each vocal

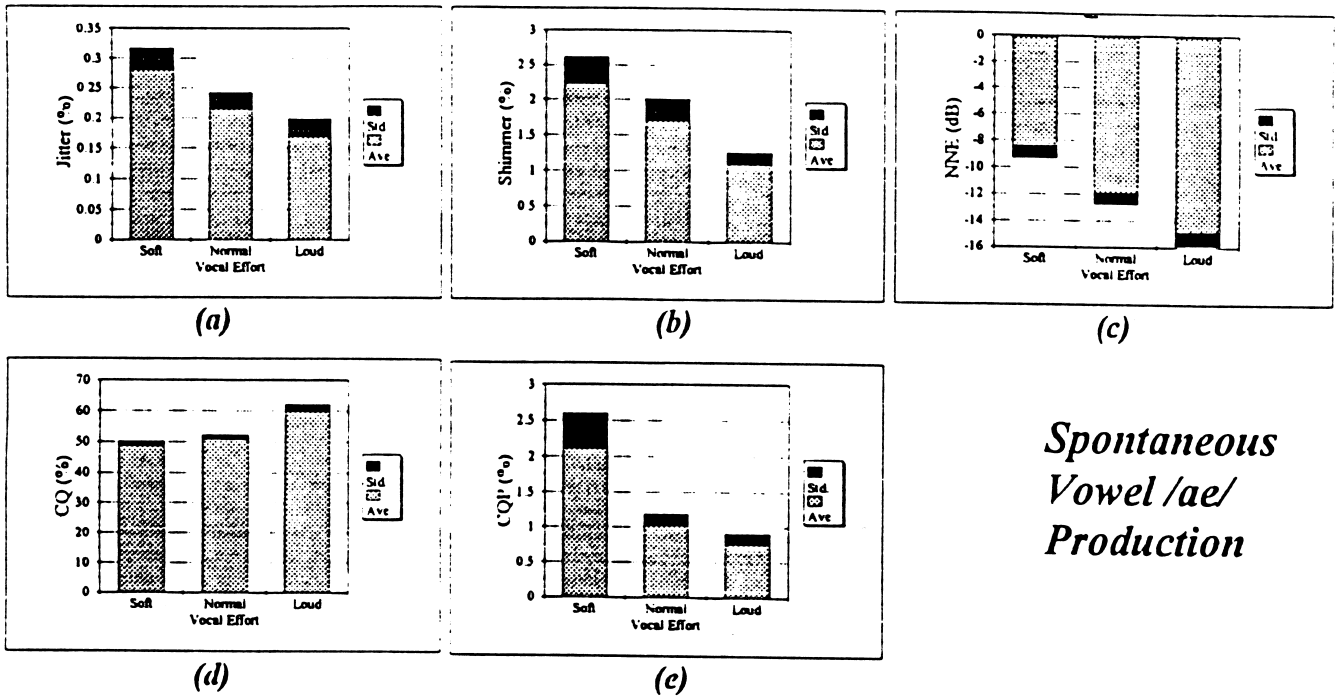


Figure 3. Jitter, Shimmer, NNE, CQ, and CQP from a sustained vowel /ae/ as a function of three vocal efforts in the spontaneous vowel production.

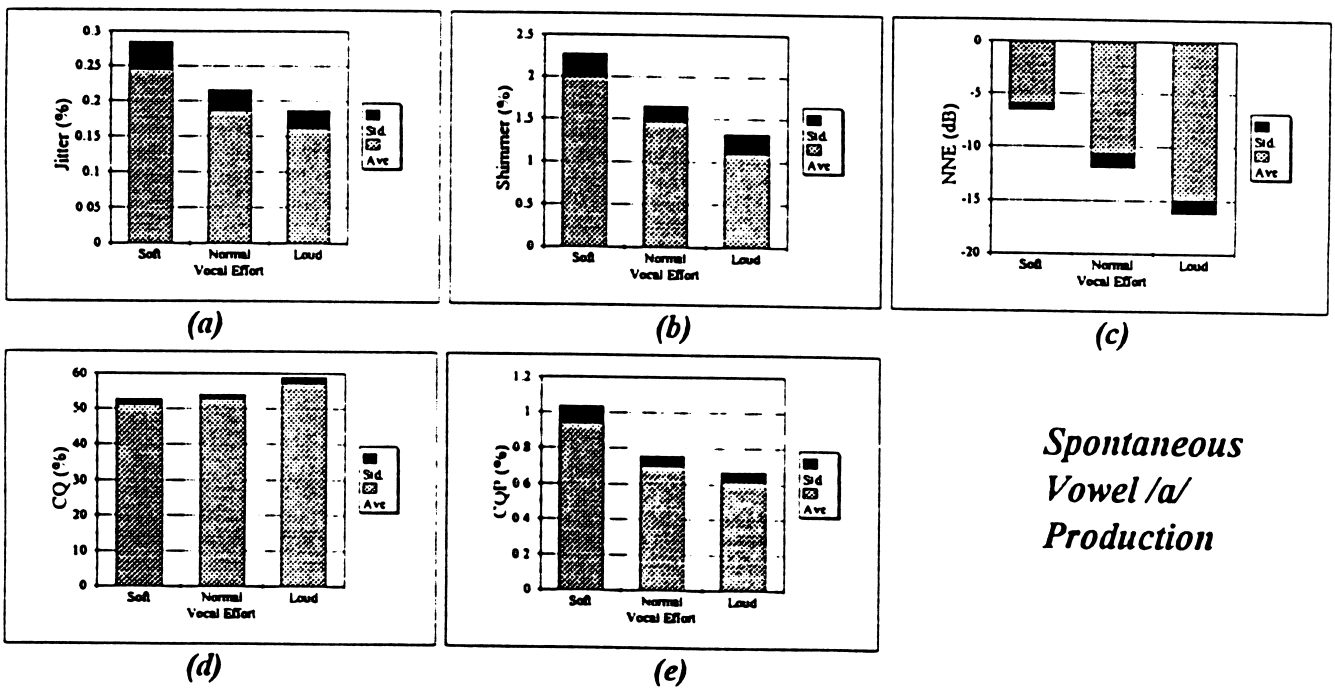


Figure 4. Jitter, Shimmer, NNE, CQ, and CQP from a sustained vowel /a/ as a function of three vocal efforts in the spontaneous vowel production.

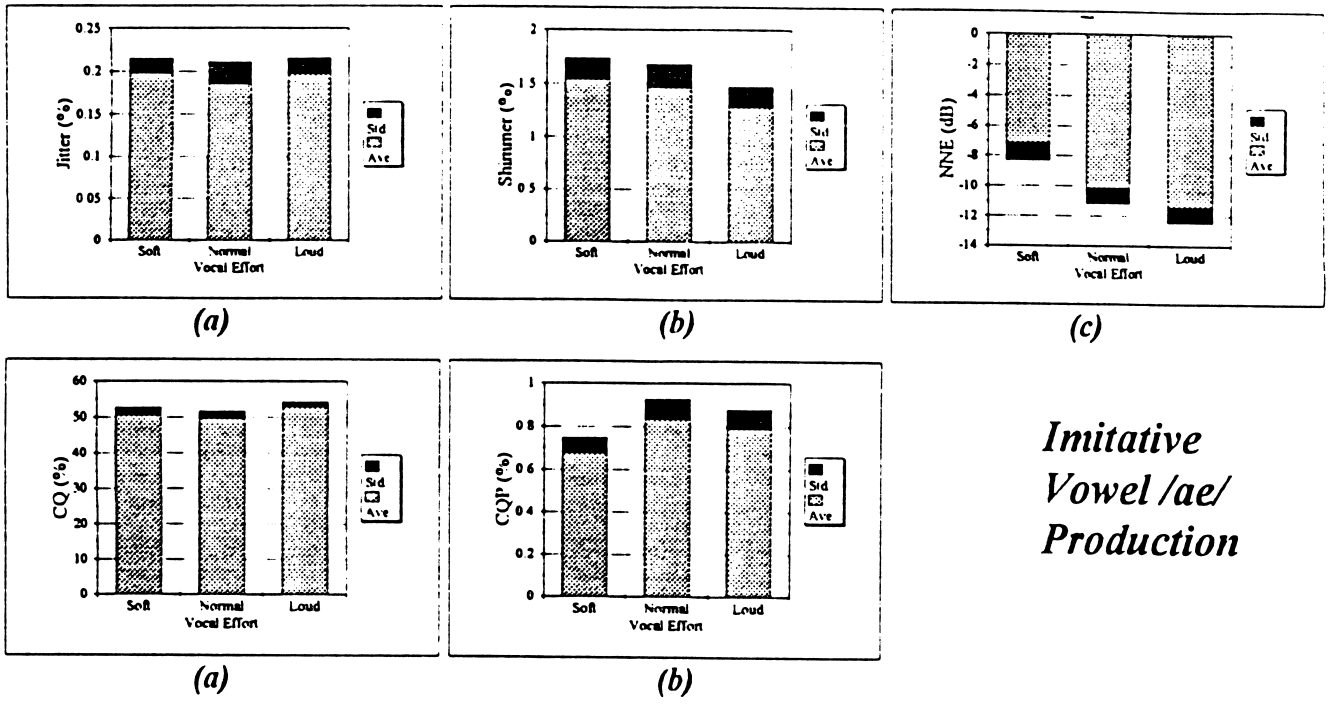


Figure 5 Jitter, Shimmer, NNE, CQ, and CQP from a sustained vowel /ae/ as a function of three vocal efforts in the imitative vowel production.

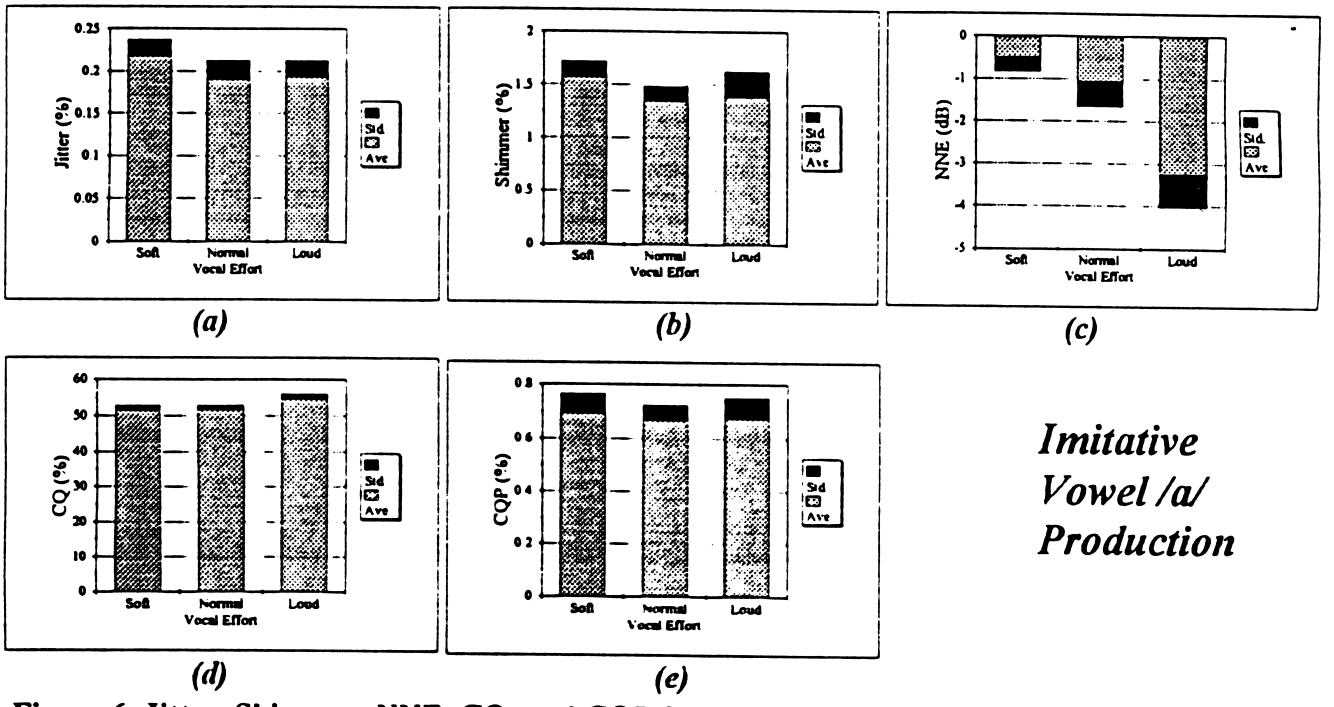


Figure 6 Jitter, Shimmer, NNE, CQ, and CQP from a sustained vowel /a/ as a function of three vocal efforts in the imitative vowel production.

cycle than during the normal and soft levels of phonation. And in the contact quotient perturbation graph (Figure 3(e)) we see that the percentage of contact quotient perturbation decreases with increasing vocal effort level.

Figure 4(a)-(e) shows the data obtained from spontaneous productions of the vowel /a/. While the values of the various measures may differ slightly from those obtained for the vowel /ae/, the patterns of change are rather similar.

TABLE 2. Statistical analysis of acoustic and electroglottographic measures at three vocal effort. The "*" indicates the significant difference at 0.05 level. (ah = /a/ natural, spontaneous, ae = /ae/ natural, spontaneous, ahm = /a/ produced in imitative of a computer stimulated /a/, aem = /ae/ produced in imitative of a computer stimulated /ae/)

	Jitter				*Shimmer*				*NNE*				*CQ*				*CQP*			
	ah	*ae*	*ahm*	*aem*	*ah*	*ae*	*ahm*	*aem*	*ah*	*ae*	*ahm*	*aem*	*ah*	*ae*	*ahm*	*aem*	*ah*	*ae*	*ahm*	*aem*
Effort Level	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Day	*	*	NS	NS	*	*	NS	*	*	*	*	*	NS	*	NS	*	NS	*	*	*
Subject	*	NS	NS	*	*	*	NS	*	NS	*	NS	*	NS	NS	*	*	*	*	NS	NS
Level @ Day	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	*	NS	*
Day @ Subject	NS	NS	NS	NS	NS	NS	*	*	NS	NS	*	*	NS	*	NS	NS	NS	NS	*	*
Level @ Subject	*	*	NS	NS	*	*	NS	NS	*	NS	NS	NS	*	NS	NS	NS	NS	NS	*	NS
L @ D @ S	NS	NS	*	NS	NS	NS	*	*	NS	NS	*	*	*	*	NS	NS	NS	*	NS	*

Figure 5 shows the vowel /ae/ produced at different vocal effort level conditions, in imitative response to target acoustic models produced by voice synthesis to reflect vocal effort levels. The target vowels for the loud, normal, and soft conditions were synthesized at 75, 72, and 68 dB SPL, respectively. Similar patterns of changes are observed in these imitative

vocalizations in comparison to those obtained during spontaneous vowel productions during changes in the vocal effort level. Figure 6 shows similar results for the /a/ vowel produced at different vocal effort levels in imitative response to the acoustic targets produced by vowel synthesis.

Shown in Table 2 are the results of numerous analyses of variance applied to the measures of jitter, shimmer, normalized noise energy, contact quotient, and contact quotient perturbation. Perhaps the most important finding from this study is related to changes occurring from changes in vocal effort level. This table shows that in all cases, changes in vocal effort level caused significant changes in the three acoustic measures and in both of EGG measures.

Discussion

In this paper, we have discussed the development of a computer program for the measurement of vocal pitch perturbation, peak-to-peak amplitude perturbation, glottal noise energy, contact quotient, and contact quotient perturbation (jitter, shimmer, NNE, CQ, and CQP), based on a newly pitch detection method. Both parabolic and bilinear interpolation methods of the AMDF provide an obvious advantage for the estimation of pitch period when compared to peak picking and zero crossing procedures. If a relatively low sampling rate is used, such as 11025 Hz, interpolations will provide an even greater advantage over these "no interpolation" procedure.

Our primary interest was to investigate the influence of vocal effort on vocal perturbation, glottal noise, CQ, and CQP measurements and to study the day-to-day variability of each influence. During our "every other day" sampling procedure for obtaining the jitter, shimmer, NNE, CQ, CQP values associated with the three vocal efforts, we observed that, in most cases, the loud vocal effort produced the lowest values. These results suggest that it is very important to control vocal effort when analyzing vocalizations, we assume that the same conclusion would apply to vocalization produced by normal and pathological subjects. On the other hand, it

appears reasonably to analyze only very steady utterances from subjects in order to get a better approximation of a speaker's typical perturbation value. This criterion may make it impossible to measure the vowel productions of some pathological speakers. As Titze (1987) has suggested, it appears best to use a voice sample at least 20-30 cycles in duration when measuring jitter and shimmer in normal speakers. Whether or not this is the case with some, or all, pathologic speakers is uncertain. What is clear is that a longer sample duration is needed in order to obtain a more stable estimate of perturbation measures. Certainly, longer vowel duration is desirable, but at a cost of increased processing time. The results of the present study suggest that more vowel repetitions are needed to determine a speaker's typical production of a given vowel. The first vowel produced during a given recording session usually yields the highest amount of variability, presumably due to psychological influence.

When tokens are recorded in a high quality digital audio tape recorder or digitized directly into a computer prior to analysis it appears to have a noticeable effect on the measures obtained.

The take-home message from this experiment is that if these acoustic and EGG measures are to be taken in the clinic, and used to compare the patient's performance from one point in time to another, it is important to have the vocalizations produced at the same vocal effort level. Secondly, Table 2 shows that in most cases there was variability in these measures from day to day. Thus, it may be important to obtain recordings from several days in order to obtain a good indication of "average" subject performance. Finally, it should be pointed out that this experiment was designed to investigate how these measures varied during vocalizations produced by normal talkers, under the prescribed conditions. It would be of considerable clinical importance to determine whether patients with voice disorders produce similar changes. Further investigations with both normal and pathologic speakers should begin to provide an answer.

Acknowledgments

We would like to thank Dr. Y. Kikuchi at Utsunomiya University and Dr. Robert Orlikoff at Memphis State University for their suggestions regarding this research project.

References

1. Baken, R.J. (1987). Clinical Measurement of Speech and Voice. A College-Hill Publication.
2. Boone, D. R. and McFarlane, S. (1988). The Voice and Voice Therapy. 4th Edition, Prentice Hall.
3. Brown, Jr. W.S, Morris, R.J. and Michael, J. F. (1989). Vocal jitter in young adult and aged female voice. Journal of Voice. 3:2:113-119.
4. Cramer, H. (1958). Mathematical Methods of Statistics. Princeton: Princeton University Press.
5. Davis, S. (1976). Computer evaluation of laryngeal pathology based on inverse filtering of speech, SCRL Monograph 13, Speech Communication Research Laboratory, Inc., Santa Barbara.
6. Fant, G. (1980). Voice source dynamics. STL-QPSR; 2-3:17-37.
7. Haji, T., Horiguchi, S., Bear, T., and Gold, W.J. (1986). Frequency and amplitude perturbation analysis of electroglottograph during sustained phonation. JASA, 80, 58-62.
8. Hirano, M., Matsushita, H., Hiki, S. (1976). Acoustic analysis for voice disorders: a basic conception for the use of acoustic measurements for the diagnosis in voice disorders. Pract. Otol. (Kyoto); 69:267-271.
9. Hirano, M. (1981). Clinical Examination of Voice. Springer-Verlag Wien New York.
10. Hirano M., (1989). "Objective evaluation of the human Voice: Clinical Aspects", Folia. Phoniatr. 41; pp. 89-144.
11. Huang, Z. and Hu, N. (1988). Research for Laryngeal Cancer Evaluation and Diagnosis", Journal of Biomechanics, Vol. 3, No.2, June, p15-p20.
12. Huang, Z. (1988). A Review of Speech Analysis and Synthesis System to Evaluate Pathological Voices, JSTU, p87-p91.

13. Huang, Z., Minifie, F. and Lin, X.. (1992). Objective Evaluation of Pathological Voices: A Preliminary Clinical Decision Program. Paper presented at ASHA, San Antonio, Texas, Nov. 1992.

14. Huang, Z., Minifie, F. and Lin, X. (1992). An Integrated Clinical Program for Voice Evaluation and Therapy. Paper presented at ASHA, San Antonia, Texas, Nov. 1992.

15. Huang, Z., Minifie, F. and Lin, X.. (1993). Measures of Vocal Function During Changes in Vocal Effort Level. Paper presented at ASHA, Anaheim, California, Nov. 1993.

16. Kasuya, H., Ogawa, S. and Kikuchi, Y. (1986) An Acoustic Analysis of Pathological Voice and Its Application to the Evaluation of Laryngeal Pathology, Speech Communication, Volume 5, No. 2, June.

17 Kasaya, H., Zue W., and Endo, Y. (1993). Measurements of Laryngeal Turbulent Noise in Pathological Voice. Paper presented at ASHA, Anaheim, California, Nov. 1993.

18. Higgins, M. B., and Saxman, J., H. (1989). A comparison of intrasubject variation across sessions of three vocal frequency perturbation indices. Journal of Acoustic Society of American. 86(3):911-916.

19. Minifie, F., Hixon, T. J. and Williams, F. (1973). Normal Aspects of Speech, Hearing, and Language. Prentice-Hall, Inc

20 Orlikoff, R. F. and Baken, R.J. (1990). Consideration of the relationship between the fundamental frequency of phonation and vocal jitter. Folia Phoniatr. 42:31-40.

21. Orlikoff, R,F and Huang, Z. (1991). Influence of Vowel Production on Acoustic and Electroglottographic Perturbation Measures. Paper presented at ASHA, Atlanta, Georgia, Nov. 1991.

22. Rothenberg, M., and Mahshie, J.J. Monitoring vocal fold abduction through vocal fold contact area. JSHR, 31, 338-351, 1988.

23.. Linville, S. E. (1988). Intraspeaker variability in fundamental frequency stability: an age-related phenomenon? Journal of Acoustic Society of American, 83(2):741-745.

24. Titze, I., Horii, Y., and Scherer, R. (1987). Some technical considerations in voice perturbation measurements. JSHR, 30:252-259.

25. Zhu, Minghui and Huang, Z. (1990) Glottal Source, Speaking and Singing Voice and Synthetic Speech of Quality. MICONEX, PRC.

High resolution spectral estimation[1]

I. Kheirallah and D. G. Jamieson

Hearing Health Care Research Unit
Department of Communicative Disorders
University of Western Ontario
Elborn College
London, Ontario, Canada

in cooperation with

AVAAZ Innovations Inc.
London, Ontario, Canada.

[1]Portions of this work were supported by grants to Dr. Jamieson from the National Sciences and Engineering Research Council of Canada and the Ontario Ministry of Health.

Abstract

This paper compares four spectral analysis techniques which are now available in commercial software packages: the short–time Fourier transform (STFT); two autoregressive methods, the autocorrelation and modified covariance methods; and a generalized time–frequency representation based on the Wigner distribution which uses a cone–shaped kernel (TFRCK). With the TFRCK, good time and frequency resolution can be obtained simultaneously. This desirable feature is not possible with any of the other three methods. In addition, when white Gaussian noise is added to the signal, it is shown that this method is able to provide an unbiased estimate of the signal without noise.

1 Introduction

Over the past decade, there have been significant advances in the methods available for the analysis of speech and other complex, time–varying signals. Several of these methods are now widely available to researchers through inexpensive software packages [4]; [9]; [6]. The choice of the analysis method has implications for the results obtained, but these implications are not obvious to researchers and clinicians who wish to analyze the characteristics of speech and other time–varying complex signals.

One of the most widely used approaches is the spectrogram, evaluated using a short–time Fourier transform. This technique has been found suitable for some applications, but it may not accurately characterize the signal under analysis. In particular, high resolution in both time and frequency is not possible.

In an effort to improve the analysis flexibility and to provide more accurate time–varying spectral estimation of speech signals, alternative techniques based on models of the speech production system have been developed. One example is autoregressive modeling of speech, where the signal is represented by an all–pole filter [5]. This quasi–stationary approach gives better results than the STFT for speech signals but is still inadequate because the signals analyzed (e.g., speech) are often nonstationary.

As a consequence of these limitations, there has been increased interest in the use of

spectral analysis methods which provide greater time–frequency resolution than conventional spectral methods. A generalized time–frequency representation where the kernel has the form of a cone has been developed in an attempt to improve the results of the previous methods [12]. In this nonstationary approach, time and frequency resolution are independent.

In this paper, the accuracy and usefulness of four spectral estimation methods are compared for synthetic signals and natural speech. The algorithms considered are the STFT, two techniques based on autoregressive modeling and the TFRCK.

2 The class of the generalized time-frequency representations

Cohen's class of generalized bilinear time–frequency representations [2] offers a unified approach to the various time–frequency analysis methods. A generalized time–frequency representation $C_x(t,f,\phi)$ of the signal $x(t)$ with kernel $\phi(t,\tau)$ is [1]

$$C_x(t,f,\phi) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi(t-t',\tau)x(t'+\frac{\tau}{2})x^*(t'-\frac{\tau}{2})e^{-j2\pi f\tau}dt'd\tau. \qquad (1)$$

This representation can be interpreted as being the Fourier transform of the convolution of the signal correlation $x(t'+\frac{\tau}{2})x^*(t'-\frac{\tau}{2})$ with the kernel $\phi(t,\tau)$. In the frequency-domain, this relation can be expressed by

$$C_x(t,f,\phi) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Phi(\xi,f-f')X(f'+\frac{\xi}{2})X^*(f'-\frac{\xi}{2})e^{j2\pi\xi t}df'd\xi, \qquad (2)$$

where $X(f)$ is the Fourier transform of $x(t)$ and $\Phi(\xi,f)$ is the two–dimensional Fourier transform of $\phi(t,\tau)$. Even though this representation is not always positive, it is considered as a time–frequency representation for specific choices of the kernel. Each of these representations is characterized by a special form of the kernel function, so that the properties of a distribution are related to the properties of its kernel. In this paper, two distributions will be analyzed – the short–time Fourier transform and a representation where the support of the kernel has the form of a cone.

2.1 The short–time Fourier transform

The short time Fourier transform belongs to the class of the generalized time–frequency representations [3]. The interpretation from the generalized time–frequency point of view allows us to find ways to try to overcome these limitations.

The spectrogram evaluated by a STFT is expressed by:

$$S_x(t,f) = \left| \int_{-\infty}^{+\infty} x(u)h(t-u)e^{-j2\pi fu}du \right|^2, \tag{3}$$

where h(t) is a real and symmetric window. The kernel of the STFT takes the following form:

$$\phi(t,\tau) = h(t+\frac{\tau}{2})h(t-\frac{\tau}{2}), \tag{4}$$

The supports of the kernel in each of t and τ directions are proportional to the duration of the window. Thus, the generalized time–frequency representation of the spectrogram shows that smoothing is introduced by the kernel in both time and frequency. The kernel of the STFT (Eq. (4)) depends on t. When it is convolved with the signal correlation (Eq. (1)), time smoothing appears. To increase time resolution, the support of the kernel in the t direction has to be decreased. To realize that, the window duration has to be reduced which will reduce the support of the kernel in the τ direction. Since the Fourier transform is performed in the τ direction, frequency resolution will decrease. Thus, increasing time resolution will decrease frequency resolution. Using a similar reasoning, it could be shown that time resolution will be reduced if frequency resolution is increased. Ideally, to overcome this tradeoff, kernels should be designed to be separable functions in t and τ, so that altering the support of the kernel in one dimension will not affect the support in the other dimension.

Although the STFT has many drawbacks, it continues to be used widely. Among the class of the generalized time–frequency representations, only the STFT always gives positive values for the power spectral density. Thus, it is able to represent the energy distribution of the signal being analyzed. In addition, the STFT is computationally efficient so that real–time implementations are possible.

2.2 Quadratic time–frequency representations

The concept of generalized time–frequency representations makes it possible to understand the reasons behind the limitations of the spectrogram and how to overcome these limitations. Eqs. (1) and (2) show that different representations can be generated by specifying different forms for the two–dimensional kernel.

To obtain good time and good frequency resolution simultaneously, an obvious choice for the kernel is one which has the narrowest possible width in time and in frequency, so that when it is convolved with the signal correlation (Eq. (1)) and the spectrum correlation (Eq. (2)), it does not introduce any smoothing. The function that has the narrowest width is the Dirac delta function. With $\phi(t,\tau) = \delta(t)$, the two–dimensional Fourier transform of this kernel is $\Phi(\xi, f) = \delta(f)$. This kernel gives the sharpest possible time and frequency resolution. The resulting time–frequency representation is the Wigner distribution.

While this representation appears to provide an ideal solution to the time–frequency resolution tradeoff of the spectrogram, it has serious limitations. When monocomponent signals are analyzed with the Wigner distribution, the results are satisfying, and frequency changes are tracked correctly without any smearing. However, with multicomponent signals, (such as speech), the Wigner distribution introduces interfering *cross terms*, artifacts which make the interpretation of the results very difficult. The presence of cross terms is due to the nonlinear nature of the Wigner distribution. To be able to use the Wigner distribution to analyze speech, further processing is therefore essential.

2.3 Time-frequency representation with cone-shaped kernel

In an effort to reduce the interference terms of the Wigner distribution while simultaneously trying to preserve its desirable properties, sophisticated smoothing functions have been developed. These functions are able to reduce the undesirable effects of the interference terms of the Wigner distribution without sacrificing its high–resolution property. A generalized time–frequency representation, based on the Wigner distribution using a cone–shaped kernel, was proposed for nonstationary signals [12]. This distribution overcomes the limitations

of the spectrogram and the artifact problem of the Wigner distribution and has been shown to provide high resolution in both time and frequency while simultaneously attenuating the interference terms [8]. The support of the kernel used has the form of a cone in the t–τ plane. Mathematically, the cone kernel is defined as [12]

$$\phi(t,\tau) = \begin{cases} g(\tau) & |\tau| \geq a\,|t| \\ 0 & \text{otherwise,} \end{cases}$$

where a is a parameter used to specify the slopes of the cone and $g(\tau)$ a tapering window.

The width of the kernel along the τ–axis specifies the frequency resolution which is inversely proportional to the length of the window T. The width of the kernel along the t–axis is independent of T which makes the time and frequency resolution independent. Interference terms are present in many generalized time–frequency representations. However, using a cone kernel, the interference terms are attenuated significantly [10].

Some specific properties of the time–frequency representation using a cone kernel are examined in [10]. In particular, to satisfy the finite time support property, a time–frequency representation should be zero whenever the signal is zero. This property is violated with most time–frequency representations when smoothing is performed in the time direction in an effort to attenuate the interference terms. An example is the spectrogram, where due to the smearing in time, the values are not always zero when the signal is zero. In contrast, this property is satisfied with the representation using a cone kernel. In addition, if the signal contains white noise, the representation using a cone kernel is able to produce an unbiased estimate of the same representation of the signal without noise [10]. In other time–frequency representations, the power spectral density of the noise is usually added to the time–frequency representation of the signal. Nonnegativity is not preserved in the cone kernel representation or in the Wigner distribution. Therefore, these representations cannot be considered as energy distributions but serve as high resolution analyses of signals in time and frequency. Thus, the representation with a cone kernel is suitable for analyzing speech signals where there is a need to resolve two closely–spaced formants, or to track a rapidly–changing spectral peak. Simultaneously, good estimates of the time of occurrence of events are possible with this technique.

3 Autoregressive spectral estimation

In an attempt to improve the resolution and spectral fidelity of the FFT, particularly for short data segments, several alternative spectral estimation techniques have been developed. These techniques represent the data to be analyzed by a model and are termed parametric methods.

Many problems associated with the FFT are attributed to the assumptions made about data falling outside the measurement interval. The finite data sequence may be viewed as a sequence of infinite length multiplied by a finite length window. The use of only these data implicitly assumes the unmeasured data are zero, which is usually not the case.

Alternative spectral estimation procedures are designed to alleviate the inherent limitations of the FFT approach. Rather than assuming that the data outside the window are zero, a more reasonable assumption is made. The process which generated the data can be modeled, and the model is then used to improve the estimate of the data falling outside the window.

Many approaches exist to determine the model parameters. Two particular cases are considered here, the autocorrelation method and the modified covariance method. Details about these two methods can be found in [7].

3.1 Weaknesses of the autoregressive techniques

As with the STFT, the autoregressive techniques use a quasi-stationary approach to analyze nonstationary signals such as speech [11]. As a consequence, the exact time of occurrence of spectral details cannot always be determined.

A serious problem with the autoregressive spectral estimation technique is its sensitivity to the addition of noise to the signal [5]. For the case of two sinusoids in noise, the resolution decreases as the signal to noise ratio decreases [5]. In addition, the spectral peaks are broadened and displaced from their true positions. This noise sensitivity occurs because the autoregressive technique uses an all-pole model to represent the signal, but this model is not correct when the signal is embedded in noise.

4 Results and discussion

A synthetic signal in clean and noisy conditions and a speech signal are used to evaluate the performance of the four algorithms discussed above, as implemented in the CSRE[2] speech analysis system [4]. The speech token was chosen to have characteristics which are traditionally hard to identify, including closely-spaced formants, rapid formant transitions, and brief components such as consonant bursts.

In each figure, the top window shows a time-frequency representation of the signal and the middle window displays the time-domain waveform. The bottom-right window shows the spectral slice at the time step where the marker in the top window is positioned and the bottom–left window displays the data used to calculate this spectral slice.

4.1 Combination of three sinusoids

To study time resolution and frequency resolution simultaneously, a signal consisting of 3 sinusoids was synthesized at a sampling frequency of 10 kHz. The first 500 ms comprised a 1 kHz tone added to a 1.25 kHz tone; the next 500 ms comprised a 3 kHz tone only (i.e., the 3 kHz tone did not overlap in time with the other two tones). For the noisy signal, a white Gaussian noise was added with 0 dB SNR.

The synthesized signal was analyzed with each of the different algorithms. Fig. 1 displays a narrowband spectrogram and Fig. 2 shows a wide band spectrogram of the same signal. The spectrogram was evaluated with a STFT. Fig. 1 shows that the 1 kHz tone and the 1.25 kHz tone are resolved in frequency but that they appear to overlap in time with the 3 kHz tone. Time resolution was better in Fig. 2 than in Fig. 1 but the separate tracks of the 1 kHz and 1.25 kHz tones are obscured (i.e., frequency resolution is much worst than in Fig. 1). Thus, the STFT may hide some characteristics of the signal (e.g., the two low-frequency tones cannot be distinguished in the wideband spectrogram) or produce others which do not truly exist (the narrowband spectrogram gives the illusion that the 3 tones overlap in time).

Analysis with the autoregressive technique using the autocorrelation method generates

[2]CSRE is a registered trade mark of Avaaz Innovations Inc.

results similar to those for the STFT. A short window gives good time resolution at the expense of good frequency resolution, while the opposite happens with a long window duration. No results were obtained for this signal with the autoregressive technique using the modified covariance algorithm, because this algorithm suffered from ill-conditioning with this signal.

A generalized time frequency distribution with a cone–shaped kernel was used to analyze the same signal, with the results displayed in Figure 3. As can be seen, this method provided good time resolution with no loss of good frequency resolution. The track of the 1 kHz tone is clearly distinguishable from that of the 1.25 kHz tone. There is no overlap in time between the low frequency tones and the 3 kHz tone. Interfering cross terms appear for a short duration during the transition but they are attenuated by about 30 dB compared to the power of signal. With multicomponent signals, the distribution using a cone kernel appears to be able to provide the excellent time–frequency resolution of the Wigner distribution, without the interfering cross terms in time or in frequency. Among the four techniques considered, this approach yields the most lucid representations of the signal. Timing information as well as the spectral content of the signal can be estimated with almost no error.

White noise at 0 dB SNR was added to the signal. The results of the autoregressive techniques were the worst, the noise heavily distorted the signal and it was not possible to distinguish the signal components. The results of the autoregressive techniques were not shown for this case, instead, the comparison is done between the best representation among the three conventional methods and the TFRCK. The results of the analysis with the STFT are shown in Fig. 4. The display is dominated by the power spectral density of the noise in the composite signal. The spectral slice centered at time 592.1 ms and displayed in the bottom right window shows the high level of the noise. Clearly, it is difficult to locate the location of the tone in this display.

The results obtained with the TFRCK are shown in Fig. 5. As can be seen, there is no noise around the tracks of the tones and the power spectral density of the noise was not added to the spectrum of the signal as occurred in the case reviewed above. This result confirms that the TFRCK can provide an unbiased estimate of the signal even in the presence of noise [10].

4.2 Speech signal

With speech signals, the results with the STFT were the worst. The results of the two autoregressive methods were very similar for the analysis window sizes used. Therefore, to be concise, the comparison was carried between the autocorrelation method and the TFRCK. The analysis parameters in each case were chosen to provide the best representation of the signal for a given method. Fig. 6 and Fig. 7 show the results of analyzing the bisyllabic /agil/ with the autocorrelation method using a short window and a long window respectively. In Fig. 6, the frequency resolution is not good, the 3rd and 4th formants of the /il/ portion are not distinguishable. In contrast, frequency resolution is improved in Fig. 7 but time resolution is reduced, the width of the /g/ is wider in Fig. 7 than in Fig. 6. In addition, in Fig. 7, the beginning of formant transitions from the /g/ to the /il/ is obscured and compared to Fig. 6, it is clearly seen that the onset of the formant transitions is shifted towards the release of the /g/.

The results with the TFRCK algorithm are shown in Fig. 8. Frequency resolution is good in this case and the tracks of the 3rd and 4th formants in the /il/ portion can be seen clearly. Simultaneously, good time resolution is preserved, the width of the /g/ and the beginning of the formant transitions are defined as well, if not better than in the previous analysis when a small window was used. For more detailed analysis, the "ZOOM IN" feature of CSRE was used to examine the portion of the displays around the /g/ release segment. This is shown in Figs. 9, 10 and 11 for the autocorrelation method with short and long analysis intervals and for the TFRCK respectively. In Fig. 9, the 2nd, 3rd and 4th formants are difficult to discern. While in Fig. 10 the formant tracks are distinguishable, the release is clearly longer in duration than in the previous case. The beginning of the release in this display is misaligned relative to the corresponding portion in the time-domain waveform shown in the middle window. Because of the poor time resolution in this case, an error of about 10 ms could be made in locating burst onset from the time-frequency display of Fig. 10. Fig. 11 shows that the TFRCK both separates the formants and simultaneously correctly estimates the duration of the release relative to the time-domain waveform. For this natural speech example, the TFRCK is therefore able to provide good time and frequency resolution

simultaneously which was not possible with the other techniques considered.

5 Conclusion

Among the spectral analysis techniques considered, the TFRCK approach was found to provide superior time–frequency resolution and appears to be better suited for the analysis of nonstationary signals. Good time resolution was obtained without degrading frequency resolution. In contrast to other spectral analysis techniques, when the signal was degraded by noise, this approach continued to provide an unbiased estimate of the signal without noise. The time–frequency representation with a cone kernel thus reveals important characteristics of speech more accurately than the other methods considered.

Figure Captions

Fig. 1: Narrowband spectrogram of the synthesized signal. It consists of 3 tones, a 1 kHz tone is added to a 1.25 kHz tone during the first 500 ms. A 3 kHz tone is generated for the next 500 ms. The low frequency tones do not overlap in time with the 3 kHz tone.

Fig. 2: Wideband spectrogram of the signal described in Fig. 1.

Fig. 3: Results of the analyzing the signal described in Fig. 1 using the TFRCK method.

Fig. 4: Spectrogram of the signal described Figure 1 in a background of Gaussian noise with SNR=0 dB.

Fig. 5: Results of processing the signal of Figure 4 with the TFRCK method.

Fig. 6: Spectral analysis of the utterance /agil/, using the autocorrelation method of autoregressive modeling. Results are shown for a short analysis window.

Fig. 7: Same analysis of Fig. 6 but using a long window.

Fig. 8: Results of analyzing the signal used in Fig. 6 using the TFRCK method.

Fig. 9: Magnification of the display in Fig. 6 showing the segment around the release burst in more detail.

Fig. 10: Magnification of the display in Fig. 7 showing the segment around the release burst in more detail.

Fig. 11: Magnification of the display in Fig. 8 showing the segment around the release burst in more detail.

References

[1] Classen, T. A. C. M. and Mecklenbrãuker, W. F. G. **1980**. "The Wigner Distribution - A Tool for Time–Frequency Signal Analysis. Part III: Relations with Other Time–Frequency Signal transformations," *Philips J. Res.* **35**, *372–389.*

[2] Cohen, L. **1989**. "Time–Frequency distributions - A Review," *Proc. IEEE*, **77**, no. 7, 941–981.

[3] Hlawatsch, F. and Boudreaux–Bartels G. F. **1992**. "Linear and Quadratic Time–Frequency Signal Representations," *IEEE Signal Processing Magazine*, **9**, no. 2, 21–67.

[4] Jamieson, D. G., Ramji, K., Kheirallah, I. and Nearey T. M. **1992**. "CSRE: A Speech Research Environment," In Ohala, J. J., Nearey, T. M., Derwing, B. L., Hodge, M.M., and Wiebe, G. E. (Editors.) *Proc. of the Second Int. Conf. on Spoken Language Processing*, 1127–1130.

[5] Kay, S. M. and Marple, S. L. Jr. **1981**. "Spectrum Analysis – A Modern Perspective," *Proc. IEEE*, **69**, no. 11, 1380–1419.

[6] Kay Elemetrics. Computerized Speech Lab. [speech analysis system]. Pine Brook, NJ.

[7] Lim J. S. and Oppenheim A. V. **1988**. *Advanced Topics in Signal Processing* (Prentice Hall, Englewood Cliffs, New Jersey), pp. 14–39.

[8] Loughlin P. J., and Pitton J. W., and Atlas, L. E. **1991**. "New Properties to Alleviate Interference in Time–Frequency Representations," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 3205–3208.

[9] Milenkovic, P. **1992**. CSpeech [computer program]. University of Wisconsin, Madison.

[10] Oh, S. and Marks, R. J. **1992**. "Some Properties of the Generalized Time–Frequency Representation with Cone–Shaped Kernel," *IEEE Trans. Signal Processing*, **40**, no. 7, 1735–1745.

[11] Riley, M. D. **1989**. *Speech Time–Frequency Representations*, (Kluwer Academic Publishers, Norwell), pp. 87–91.

[12] Zaho, Y., Atlas, L. E., and Marks R. J. **1990**. "The Use of Cone–Shaped Kernels for Generalized Time–Frequency Representations of Nonstationary Signals," *IEEE Trans. Acoust., Speech, Signal Processing*, **38**, no. 7, 1084–1091.

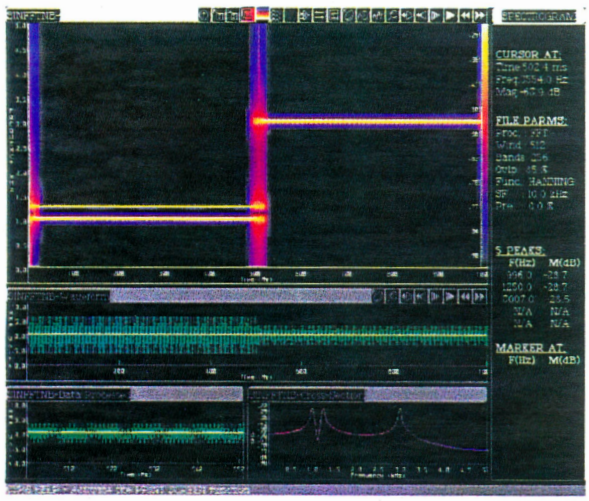


Fig. 1

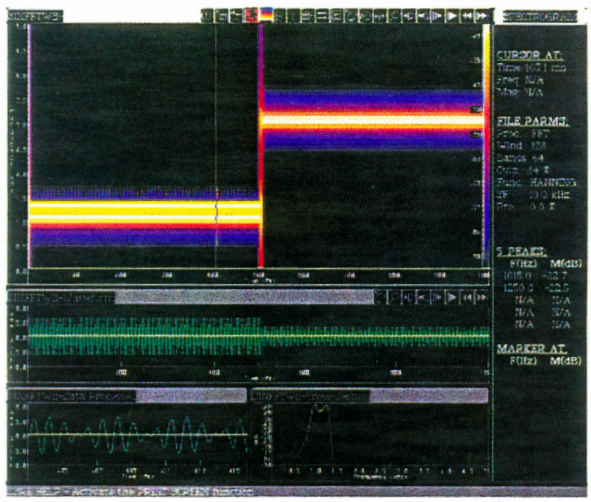


Fig. 2

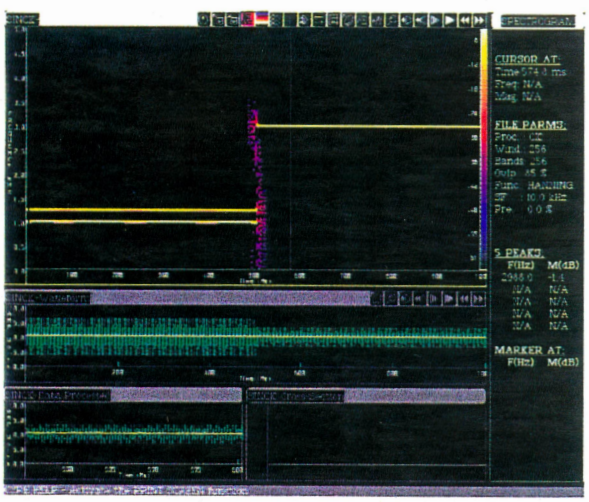


Fig. 3

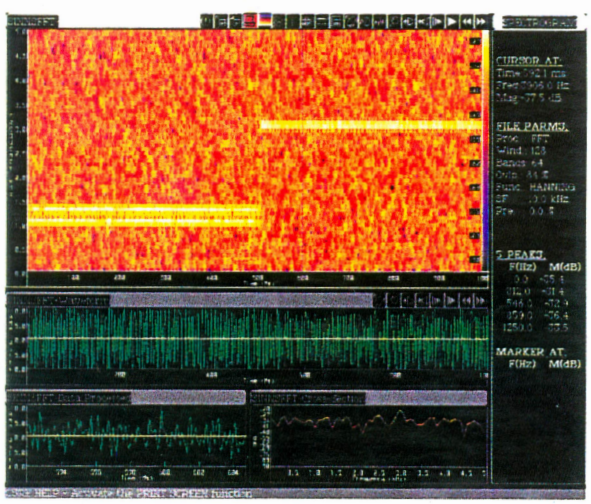


Fig. 4

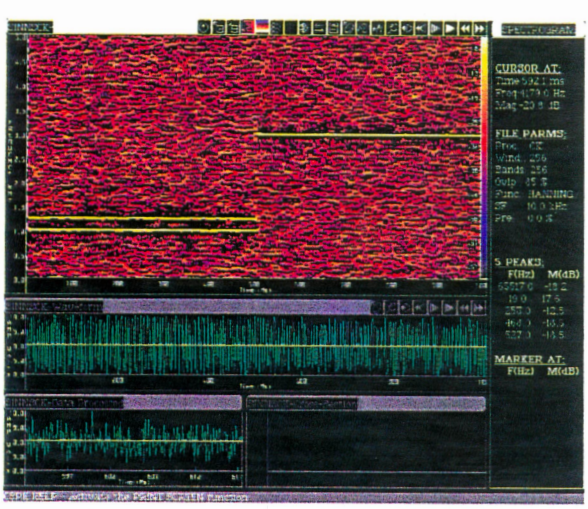


Fig. 5

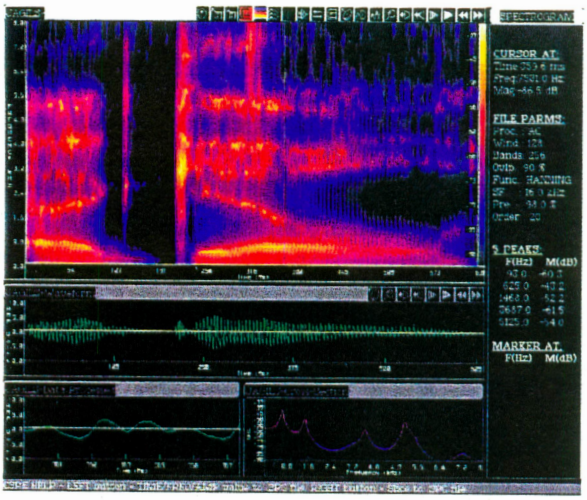


Fig. 6

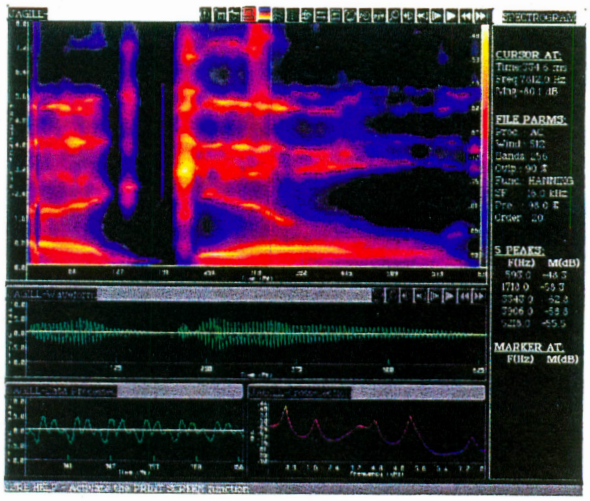


Fig. 7

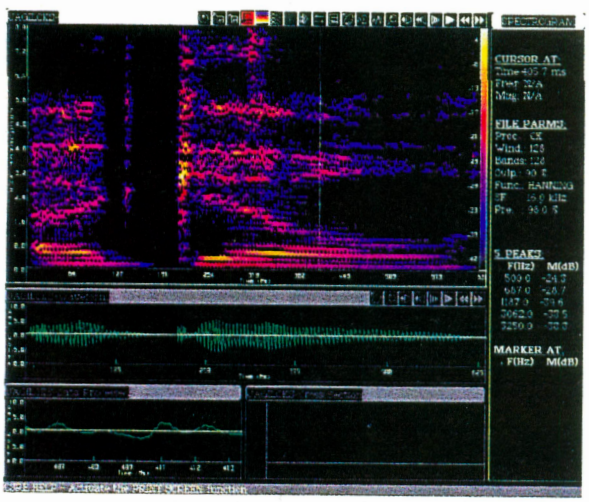


Fig. 8

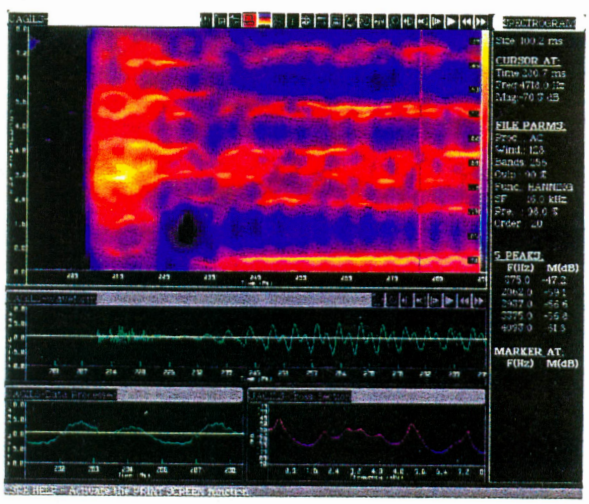


Fig. 9

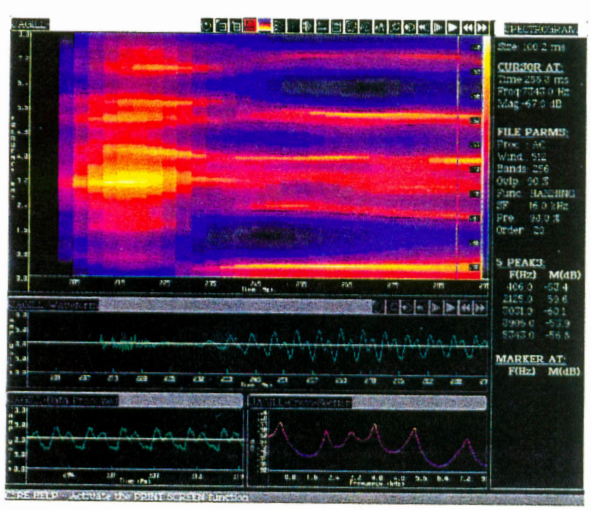


Fig. 10

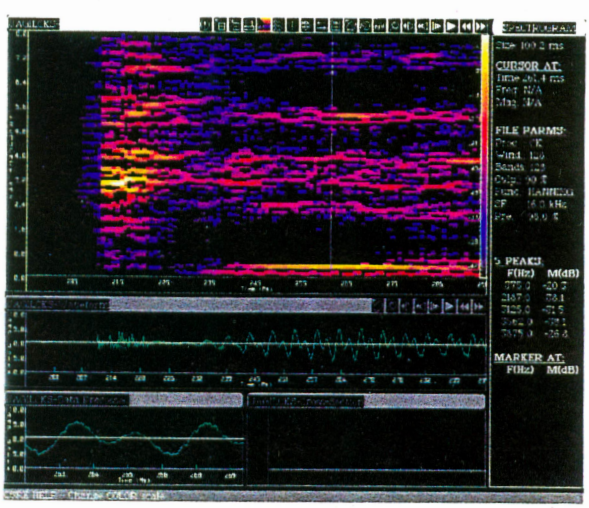


Fig. 11

MECHANISMS OF JITTER-INDUCED SHIMMER IN A DRIVEN MODEL OF VOCAL FOLD VIBRATION[1]

Darrell Wong, Robert Lange, Ingo R. Titze[2], Chwen Geng Guo

National Center for Voice and Speech
and
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
[2]Department of Speech Pathology and Audiology, University of Iowa

INTRODUCTION

Measurements of jitter and shimmer on voice signals are obtained for the purpose of discerning perturbations in the oscillatory behaviour of the vocal folds. Jitter is defined as the average cycle-to-cycle change in the fundamental period length, and shimmer is the average cycle-to-cycle change in amplitude, but there is no generally accepted definition of amplitude for a complex signal. Traditional definitions rely on specific events in time, such as the largest positive peak-to largest negative peak, or the largest negative peak value. Other definitions include the root-mean-squared (RMS) value of each cycle (given that the fundamental period markers have been correctly placed) used by Kempster and Kistler (1984) and Hillenbrand (1987), or the gain factor defined by Milenkovic (1987).

All of these amplitude definition and extraction algorithms work well with amplitude modulated (AM) voicing signals, reporting linear increases in shimmer as the extent of the AM is increased. Their behaviour under frequency modulation (FM), however, is nonlinear. This nonlinearity is often attributed to time-aliasing (Qi, 1994); (Oppenheim, 1989). The phenomenon has been observed by both Hillenbrand and Milenkovic and found to vary with vowel and F_0. The concept of time-aliasing stems from the impulsive nature of the source-filter model of speech. The vocal tract is modeled as a linear filter excited by an impulse train. From cycle to cycle, preceding vocal tract impulse responses overlap and add to the current cycle. When the train is aperiodic, the changing phase relationships between consecutive impulse responses cause the signal to become distorted relative to previous cycles, resulting in measurable shimmer. Since all of the amplitude definitions defined above extract measurements at a stage where the time-aliasing has already occurred, they all suffer from its effects.

As a consequence, whenever jitter is present, any measurement of shimmer cannot be solely attributed to amplitude modulation of the vocal fold oscillations. It would thus be useful to

[1]This work has been funded by the National Institutes of Health, Grant DC 00387-08

devise techniques that can ignore or overcome aperiodic time aliasing as a shimmer source, or else identify and analyze other signals that more accurately reflect the fold behaviour.

In this study the second alternative is pursued. We examined the behaviour of data obtained from a time domain simulation model that generates signals at the level of the vocal folds and throughout the vocal tract. This enabled us to characterize possible sources of shimmer induced by FM modulation of a driven model of vocal fold tissue displacement. It was found that vocal tract time aliasing is only one of a number of sources of shimmer originating from different mechanisms and affecting various signals. Shimmer also appears to be a result of the nonlinear interactions between the transglottal acoustic pressure, mucosal wave velocity, and tissue displacement. These jitter-induced amplitude perturbations then propagate to the supraglottal pressure P_{in}, the subglottal pressure P_s, the minimum glottal area A_g, the glottal flow U and its time derivative (dU in this paper), and finally the output pressure from the vocal tract P_0. In this paper, we study the possible mechanisms for these perturbations.

METHOD AND MODEL

An interactive computer simulation of the vocal fold and vocal tract system has been used to model the behaviour of the folds under conditions of FM subharmonic modulation of the vocal fold tissue displacement . Subharmonic modulation of order 1/2 was chosen so that any changes in the plots could be observed by inspection. An FM extent level of 30% was chosen to exaggerate the effects, although it is acknowledged that this is much greater than values typically found in human phonation. The model, SPEAK-model 2, was developed at the University of Iowa by one of the authors. It incorporates source-tract interaction, but does not incorporate self oscillation. Instead, there is direct control of tissue displacement by a mathematical driving function. Empirical relationships previously described in the literature involving F_0, vocal fold length and thickness, mucosal wave velocity, lung pressure, and transglottal pressure are used to define the behaviour of the system. The effects of modulation are demonstrated as mechanisms supported by these equations. Figures from the simulations are used to illustrate the phenomena.

Modeling equations

A full description of the model appears in (Titze 1995). A brief outline is given here. Consider the top view of the vocal folds (Figure 1). The vocal processes are at $x = +/- \zeta_0$, at $y = 0$. The lowest mode of vibration occurs in the y-direction (a half sinusoid). When ζ_0 is small, complete closure can occur, while for large values, a glottal chink causing flow leakage occurs (the parameter h in Figure 1 indicates the height of the chink).

For the simple case without jitter, the edges of the vocal folds oscillate sinusoidally in time with an angular frequency $\omega = 2\pi F_0$. The glottal width at any point on the y axis (as the folds

move outward) is defined as

$$g(y,t) = 2[\zeta_0(1-y/L) + \zeta_m \sin(\omega t)\sin(\pi y/L)] \quad (1)$$

where L is the vocal fold length, and ζ_m is the maximum vocal fold displacement (at $y = L/2$). The first term in (1) is the prephonatory initial displacement.

To make the model respond in a manner similar to the human folds, relationships extracted from the literature have been utilized. For example, it is known that ζ_m is dependent on lung pressure and F_0. The following rule is adopted from Titze (1988):

$$\zeta_m = 17.4 P_L^{0.5} F_0^{-1.6} \quad (2)$$

where P_L is the lung pressure.

Figure 2 shows a three dimensional view of the folds. A z-axis has been added to describe the motion of the mucosal wave as it travels from the bottom to the top of the folds. If the vertical dimension is sliced into layers, each layer k can have its own value of initial condition ζ_{0k} and maximum displacement ζ_{mk}. The propagation of the mucosal surface wave can then be obtained by replacing $\sin(\omega t)$ in equation (1) with $\sin(\omega(t-z/c))$, where z is the vertical point under consideration and c is the wave velocity. If the vertical axis is sliced into N layers, the phase between layers is assumed to vary linearly between layers from the bottom (k=1) to the top (k=N) of the z-axis:

$$\phi_k = \omega(t-z/c) = \omega(t-kT'/(Nc)) \quad (3)$$

where ϕ_k replaces ωt in equation (1) and T' is the thickness from bottom to top. Titze, Jiang and Hsiao (1993) conducted an experiment to measure c. They examined the motion of two sutured points on a canine hemi-larynx placed in the vertical dimension. The time required for the displacement of the superior suture to 'catch up' to the displacement of the inferior suture was measured by observing the time taken to pass a certain point. This time was interpreted as an indication of the wave propagation speed. In the results section of the paper, the authors discuss the fact that the inter-suture distance (and the overall thickness of the folds) may vary during vibration as a function of F_0.

We present an equation which takes this thickness variation into account:

$$c/T' = aF_0/T_0 \quad (4)$$

where T_0 is the nominal thickness at rest, T' is the thickness during vibration for a given F_0, and *a* is a constant parameter of value 0.01 meters (Titze, 1995).

From the above driving equations, the displacement x at any point (y,k) may be obtained:

$$g(y,k,t) = 2[\zeta_{0k}(1-y/L) + \zeta_{mk}\sin(\phi_k)\sin(\pi y/L)] \quad (5)$$

The glottal area between the symmetric folds at any layer k can be calculated by integration, and the minimum glottal area A_g, observable by a photoglottogram, can be estimated by finding the minimum area layer at each point in time.

The glottal flow is determined by the transglottal pressure. The orifice A_g, the subglottal pressures P_s, and the supraglottal pressure P_{in} determines the flow U via the equation:

$$P_s - P_{in} = k_t \rho |U| U / A_g^2 \quad (6)$$

where k_t is the transglottal pressure coefficient (assumed constant for this study), and ρ is air density. An /a/ vocal tract shape is modeled using wave reflection equations (Liljencrants, 1985).

FREQUENCY MODULATION

If F_0 is sinusoidally modulated, then the 'instantaneous fundamental frequency' can be written as:

$$F_0' = F_0(1 + E\cos(2\pi F_m t)) \quad (7)$$

g(y,t) in equation (1) thus becomes frequency modulated at a modulation frequency F_m and an extent E, proportional to F_m/F_0. Figure 3 illustrates the basic shape of the modulated displacement when F_0 is 125 Hz, F_m is 62.5 Hz, and E is $0.2(2\pi F_m/F_0)$. The effects of Equation 7 propagate through the model in the following ways.

Modulation of the Mucosal Wave Velocity

If Equation 7 is assumed, the mucosal wave velocity c now varies with F_0' (via Equation 4). The modulation effect then propagates into Equations 3 and 5. Before we discuss what happens in our model, let us examine what happens in a simpler system of equations. Consider the signal in Figure 3. It might represent the oscillation of the bottom layer (k=1) of the folds. If a constant *time* delayed version of this signal is used to represent the top layer (k=N) of the vocal folds, amplitude variations will occur in the minimum glottal area wave A_g. These variations in the minimum glottal area are depicted in Figure 4. Such peak-to-peak variations in the 'minimum aperture' appear regardless of the relative sizes of the waves. However, consider the top of the folds to be oscillating

with a constant *phase* delay relationship. The minimum glottal area exhibits no amplitude variations here (Figure 5). It can be shown that the minimum glottal area in Figure 5 will always be free of amplitude variations, regardless of the relative sizes of the top and bottom displacements. This is provided that a constant phase lag is maintained between the layers and that each of the contributing displacements exhibit no amplitude variations independently.

In our vocal fold model, if c is proportional to the 'nominal' average value F_0, it can be shown analytically that a constant time delay mucosal wave will be observed (see Appendix). However, if c is instead proportional to the instantaneous value of fundamental frequency F_0', it is also shown in the Appendix that a constant phase delay relationship will be produced. Studies on the mucosal wave e.g. (Titze 1989) report a constant phase delay, although an FM situation is not usually considered. It remains to be demonstrated which type of delay - constant time, constant phase, or neither - actually occurs for FM modulations of the vocal folds.

Figure 6 demonstrates what happens in the driven vocal fold model when the nominal average value of F_0 is used for both the maximum vocal fold displacement and the speed of mucosal wave propagation. The displacement signal x exhibits FM behaviour, but no amplitude perturbation. The glottal area waveforms (nonminimum) for three points on the z axis (bottom (k=1), middle (k=10), and top (k=21)) show a similar result to Figure 4. The FM subharmonic in x produces amplitude perturbations in minimum glottal area A_g, the flow U, the derivative with respect to time dU and on into other signals in the vocal tract. Note that the contact area CA does not exhibit any amplitude perturbations, and that the x and A_{g1}, A_{g10}, and A_{g21} plots are on an expanded time scale relative to the other plots, so that the effect of overlapping A_g waves can be seen. This applies to Figures 6 through 10.

Figure 7 demonstrates the situation when the instantaneous F_0' is used for the mucosal wave speed. A constant maximum amplitude is assumed (time varying ζ_m will be discussed later). The three glottal area waves exhibit no visible minimum aperture amplitude variation in A_g or CA, as predicted by the previous discussion. The glottal flow U, however, does exhibit maximum amplitude variations, which then propagate into dU and on into the vocal tract. The cause of this variation is discussed next.

Transglottal Pressure and Ag,U slopes

The glottal flow in Figure 7 demonstrates amplitude perturbation even though A_g does not. This occurs because of a combination of two effects. First, from Equation 6, the glottal flow U is related to A_g and the transglottal pressure P_s-P_{in}. P_{in} is related to vocal tract loading, which is usually inductive, causing the flow to skew to the right (i.e. it is delayed relative to A_g). As a result, the peak of the flow wave U occurs after the peak of Ag. While the peaks of Ag are the same height, the negative closing slope varies due to the FM modulation. As a result, the value of

A_g at the instant of peak flow will vary, causing changes in U.

This phenomena can be more clearly seen in Figure 8, in which the model is modulated with a 1/3 FM subharmonic. Again, A_g exhibits no amplitude perturbation, while U, dU and subsequent propagated signals exhibit amplitude changes. The slope of A_g clearly influences U, which in turn influences dU. The negative slope in both U and A_g appears to be the primary determinants of peak magnitudes in the vocal tract. It should be noted that the negative minimum peak in dU is often associated with the excitation of the vocal tract, where the initial large negative peak in the voice signal is related to this peak in dU.

Modulation of ζ_m

If ζ_m were constant, the glottal width equation (1) would vary sinusoidally for constant F_0, and in a modulated manner (Figure 3) for F_0'. When a time varying relationship between ζ_m and F_0' is assumed for Equation 2 (again assuming that these equations apply to dynamically modulated conditions), direct amplitude modulation of the maximum tissue displacement occurs. Figure 9 assumes a constant phase delay (F_0' in Equation 4) and dynamically varying ζ_m (F_0' in Equation 2). The displacement x shows both amplitude and frequency modulation. As a result, amplitude perturbation appears in all subsequent signals (except for CA).

DISCUSSION

This study has identified two potential laryngeal sources for jitter-induced shimmmer, and suggested that caution be used regarding the interpretation of another. One source is the modulation of the maximum amplitude of vibration (ζ_m), which directly influences the amplitude of the glottal areas. The second source is the slope of the minimum glottal area, which determines the peak in the flow wave (due to the inductive load of the vocal tract). It should also be noted that it is the slope of the flow wave, not the peak, which is closely tied to the impulsive excitation of the vocal tract pressure wave. Often it is this peak in P_0 that is marked in voice waveform analysis.

If mucosal wave velocity c is dependent on F_0', it has been demonstrated here that a constant phase delay rather than a constant time delay occurs between the top and bottom of the folds, making it unlikely (in this model of vibration) that minimum glottal area (A_g) amplitude perturbation due to a constant time delay mucosal wave is a source of shimmer.

It should be remembered, however, that the F_0' and ζ_m modulations likely occur simultaneously, resulting in A_g perturbation as illustrated in Figure 9. In a real vocal fold, where changes in the stiffness of the muscle are likely to be the source of perturbation, it is probable that both F_0' and ζ_m will vary.

It is interesting to note that the contact area CA is amplitude insensitive to FM. CA can be measured with the electroglottograph and it would be useful to know whether it demonstrates

shimmer or not. It is a measure of the electrical conductivity measured from one side of the folds to the other. In all the examples that were given, complete closure was achieved (the z layers were preconfigured so that the initial ζ_{0k} were close together, causing CA to reach a maximum for all cycles. Figure 10 illustrates the case where this assumption is relaxed, resulting in incomplete closure for some of the layers in the z-axis. As a result, CA exhibits amplitude perturbation. It should thus be noted that the ability of CA to demonstrate glottal displacement amplitude perturbation is limited due to 'saturation', because it is inherently a measure of behaviour during collision rather than glottal opening.

Since the electroglottograph does not directly reflect maximum tissue displacement, we should ask which signal is likely to most accurately describe the tissue displacement. Assuming that the vocal fold mucosal wave is a constant phase delay phenomena, then the glottal area signal will directly describe the displacement. This is best measured by the photoglottograph.

SUMMARY

The interactions among several voice production variables have been qualitatively described for a driven model of vocal fold displacement. The mechanisms identified here suggest that shimmer in the glottal flow U occurs from slope changes in A_g, while direct modulation of the peak tissue displacement ζ_m will affect A_g.

The observations made here suggest further study. The model we have studied drives the tissue displacement explicitly using predetermined fundamental frequency behaviour. While the mechanism due to slope changes in A_g is likely to be common to all vocal fold models, the empirical equations relating ζ_m and c to F_0 are not used in self-oscillating models, since F_0 is not directly controlled. An examination of self-oscillating models in which physiological variables are controlled (e.g. tissue stiffness) is warranted.

APPENDIX

Consider the tissue displacement equations ζ_t and ζ_b for the top and bottom of the fold:

$$\zeta_b = \zeta_{0b} + \sin(\omega_0' t) \qquad \text{A1}$$

$$\zeta_t = \zeta_{0t} + \sin(\omega_0'(t - T'/c)) \qquad \text{A2}$$

where ζ_{0b} and ζ_{0t} are the initial displacements, ω_0' is the fundamental frequency (possibly time varying), t is the time variable, T' is the variable vertical thickness, of the fold and c is the velocity of the traveling wave.

If F_0 is assumed constant, $c = aF_0T'/T_0$, then a constant time delay equation replaces A2:

$$\zeta_t = \zeta_{0t} + \sin(\omega_0'(t - T_0/aF_0)) \qquad \text{A3}$$

On the other hand, if F_0 (and therefore c) is assumed time varying, then $c = aF_0'T'/T_0$ or $c = a\omega_0'T'/2\pi T_0$, and A2 becomes a constant phase lag equation:

$$\zeta_t = \zeta_{0t} + \sin(\omega_0't - T'\omega_0'2\pi T_0/(T'a\omega_0')) \quad \text{A4}$$

$$= \zeta_{0t} + \sin(\omega_0't - 2\pi T_0/a) \quad \text{A5}$$

BIBLIOGRAPHY

Hillenbrand, J., "A Methodological Study of Perturbation and Additive Noise in Synthetically Generated Signals". *Journal of Speech and Hearing Research*, 30 (4), 448-461, December 1987.

Kempster, G.B., and Kistler, D.J., "Perceptual Dimensions of Dysphonic Voices". *Journal of the Acoustical Society of America*, 75 (Suppl. 1) S8 (A), 1984.

Liljencrants, J., "Dynamic Line Analogs for Speech Synthesis". *Quarterly Progress and Status Report, STL-QPSR*, 1/1985, 1-14. Speech Transmission Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden.

Milenkovic, P., "Least Mean Square Measures of Voice Perturbation", *Journal of Speech and Hearing Research*, 30 (4), 529-538, December 1987.

Oppenheim, A., and Schafer, R., *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

Qi, Y.Y., Weinberg, B., Bi, N., and Hess, W.J., "Minimizing the Effect of Period Determination on the Computation of Amplitude Perturbation in Voice", *NCVS Acoustic Voice Analysis Workshop Proceedings*, Denver, Colorado, Jan. 17-18, 1994.

Titze, I.R., *The Myoelastic-Aerodynamic Theory of Phonation*, in progress (1995).

Titze, I.R., "The Physics of Small Amplitude Oscillation of the Vocal Folds". *Journal of the Acoustical Society of America*, 83 (4), 1536-1552, 1988.

Titze, I.R., Jiang, J., and Hsiao, T.Y., "Measurement of Mucosal Wave Propagation and Vertical Phase Difference in Vocal Fold Vibration", *Annals of Otol. Rhinol. Laryngol.* 102, 58-63, 1993.

Titze, I.R., "On the Relation Between the Subglottal Pressure and Fundamental Frequency in Phonation". *Journal of the Acoustical Society of America*, 85 (2), 901-906, 1989.

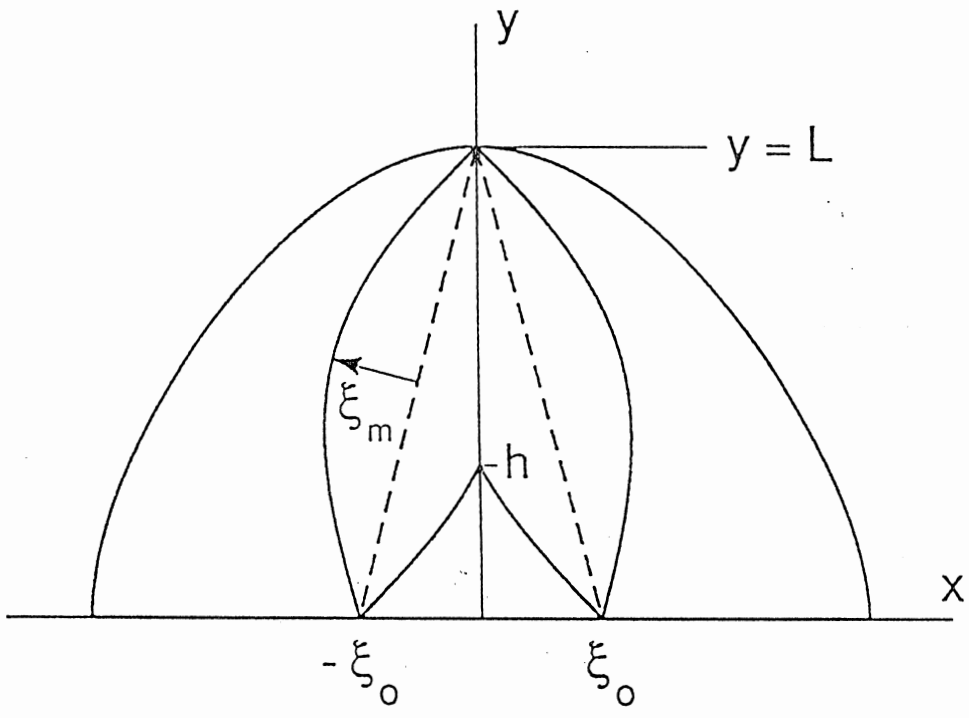


Figure 1. Top view of vocal fold model. From (Titze 1994).

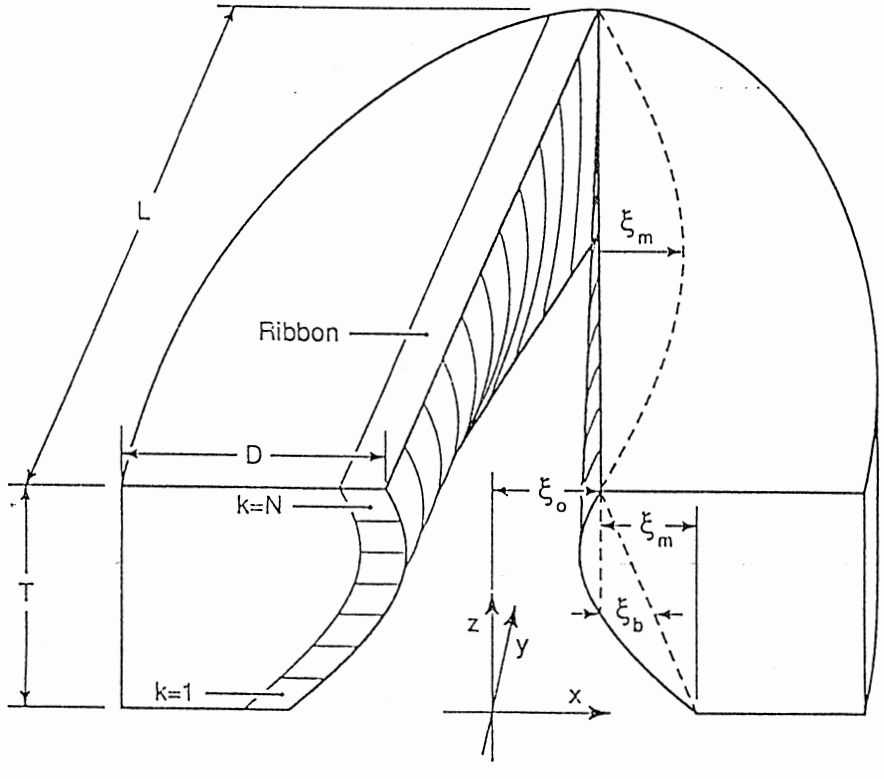


Figure 2. Three dimensional view of vocal folds. From (Titze 1994).

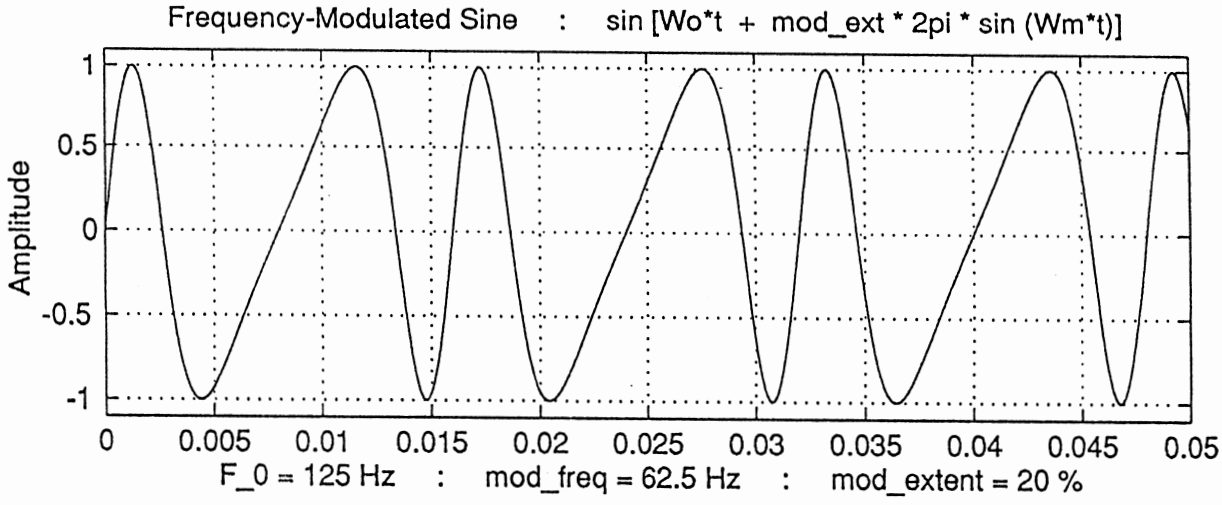


Figure 3. Subharmonic Frequency Modulated Sinusoid. F_0 = 125 Hz, F_m = 62.5 Hz, E = 20%

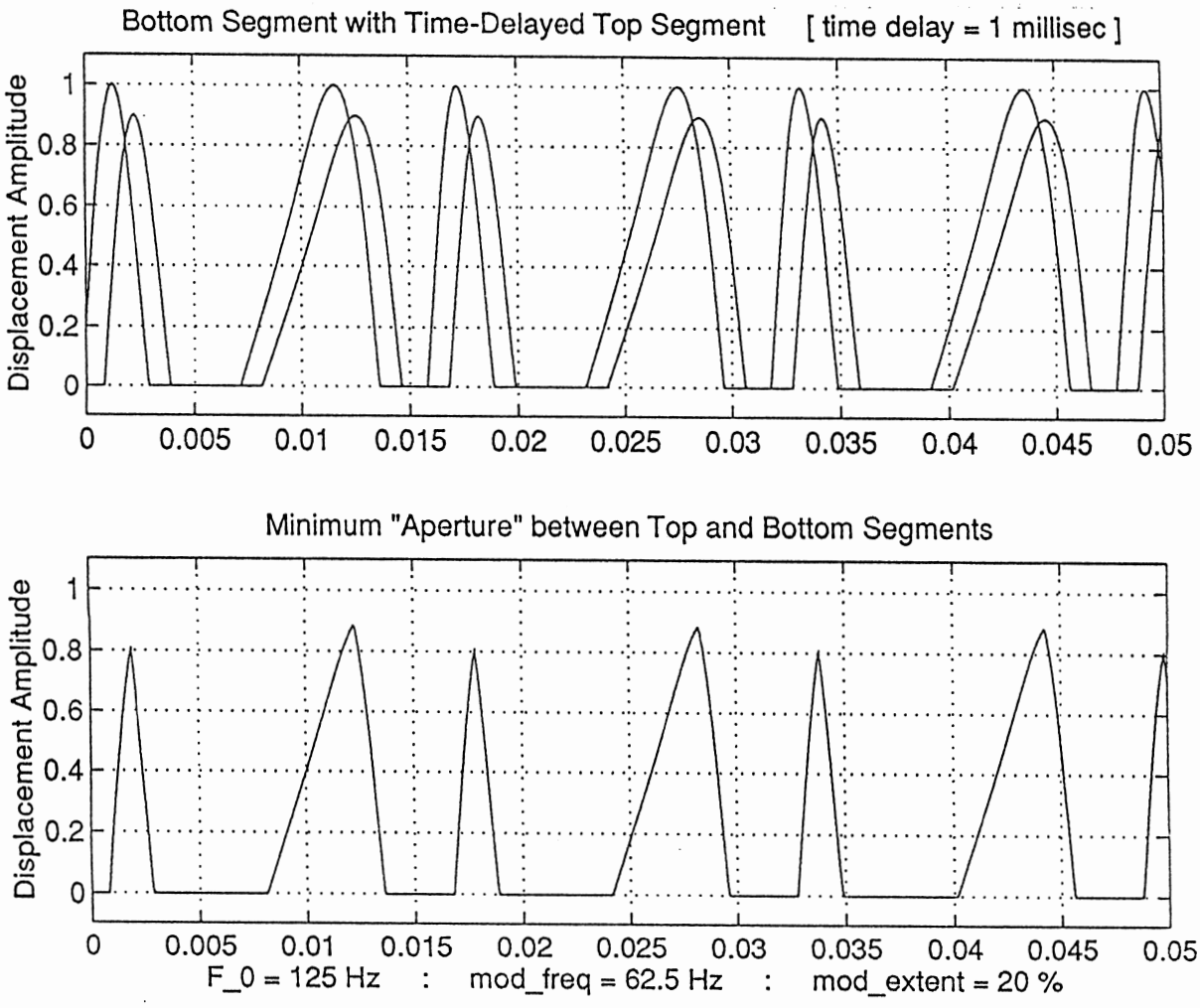


Figure 4. Time delayed FM subharmonic (1/2) modulated sine waves and the minimum aperture wave result.

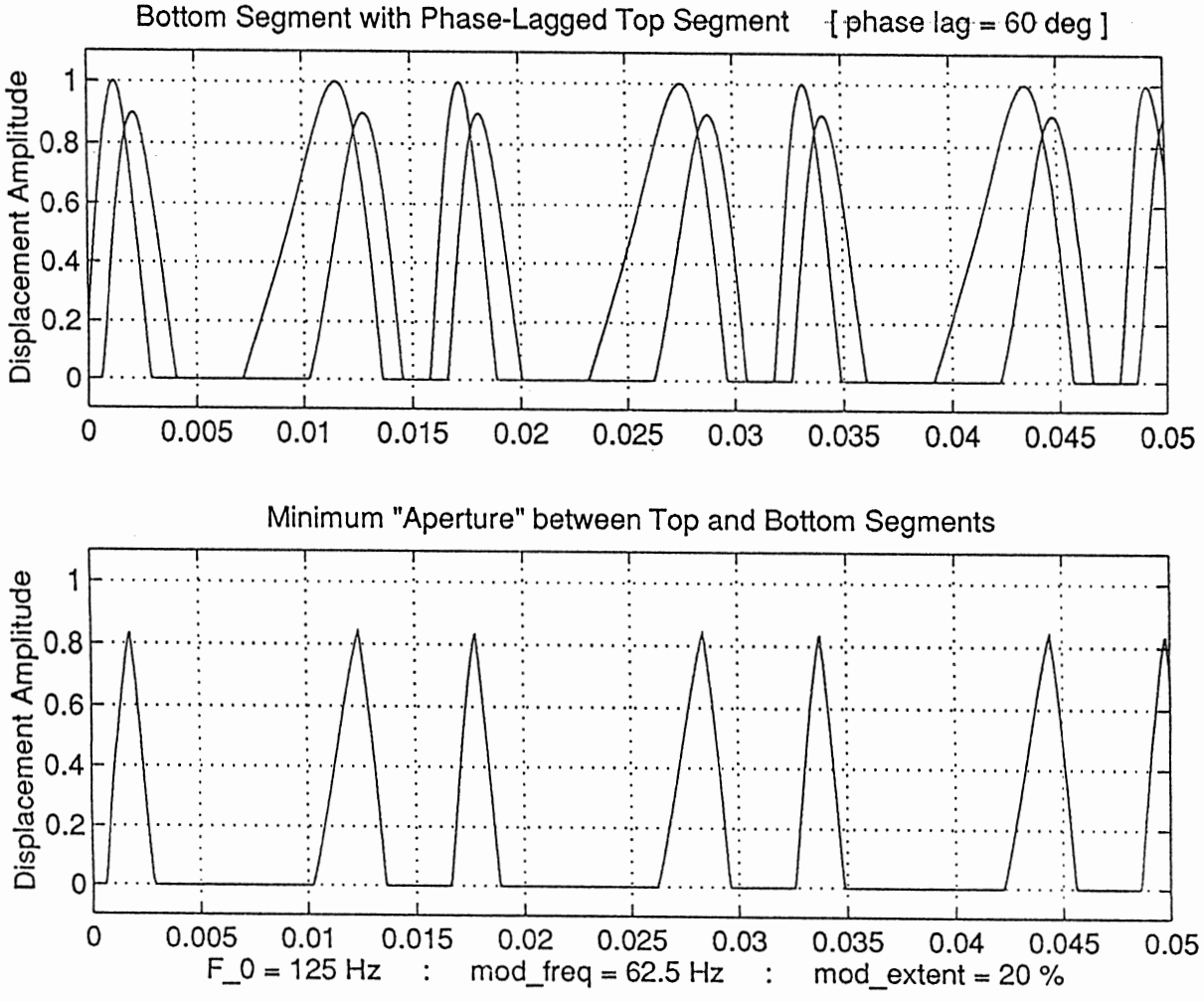


Figure 5. Phase delayed FM subharmonic (1/2) modulated sine waves and the minimum aperture wave result.

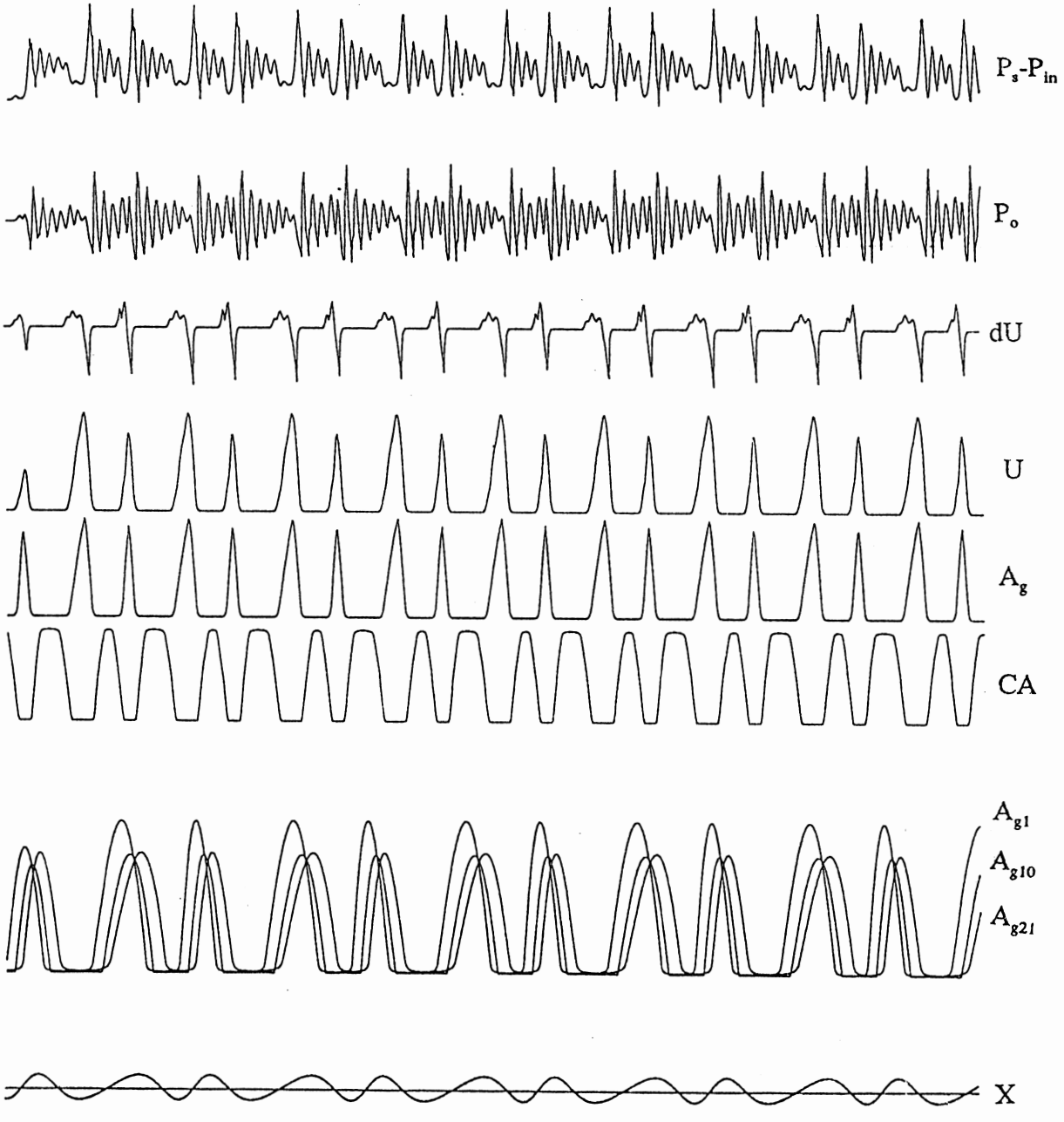


Figure 6. SPEAK generated subharmonic (1/2) array of signals. Assumes F_0 is constant for Equations 2 and 4, causing a constant time delay between displacements.

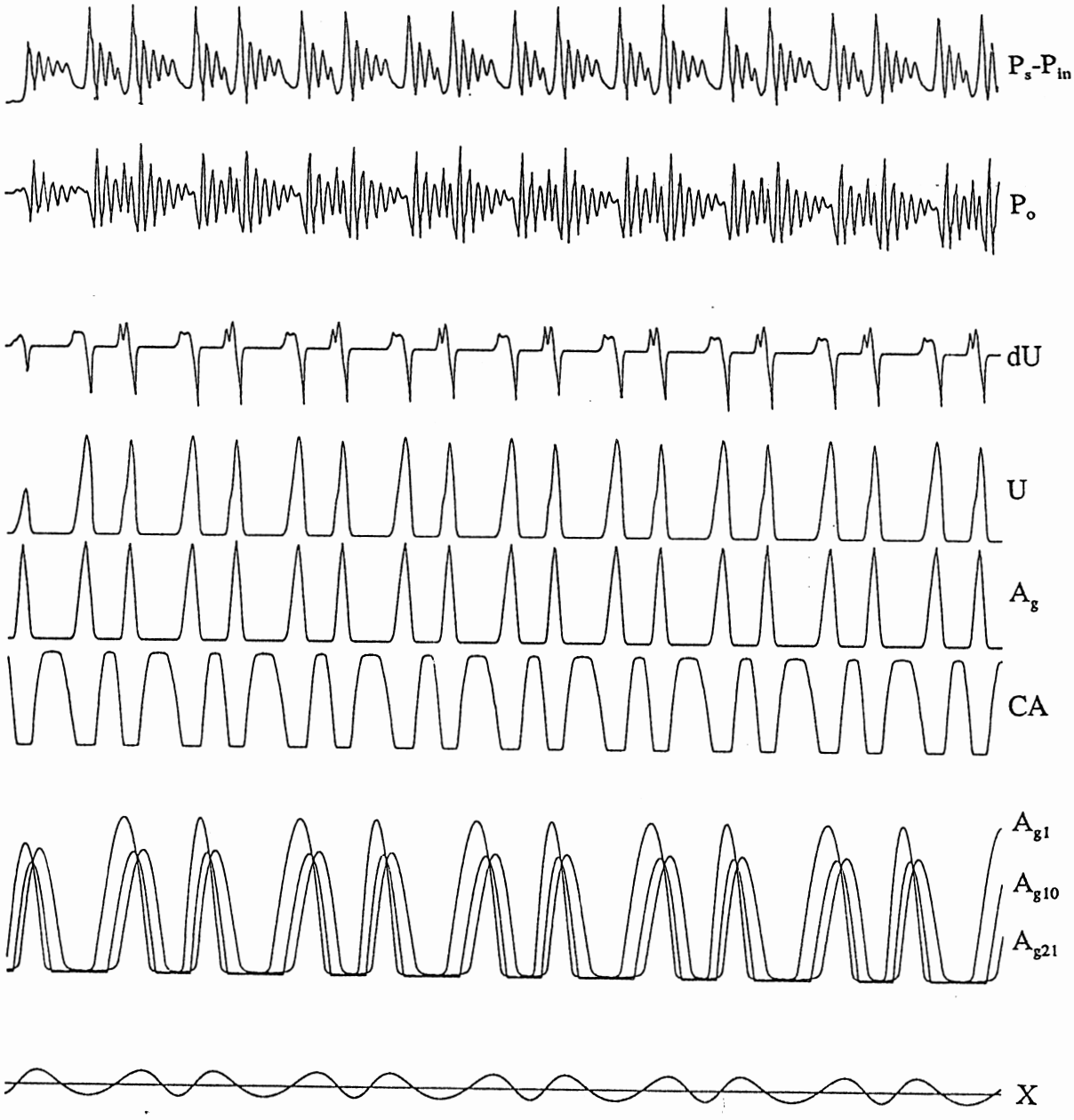


Figure 7. SPEAK generated subharmonic (1/2) array of signals. Assumes F_0 is constant for Equation 2, and varies for Equation 4, causing a constant phase delay between displacements.

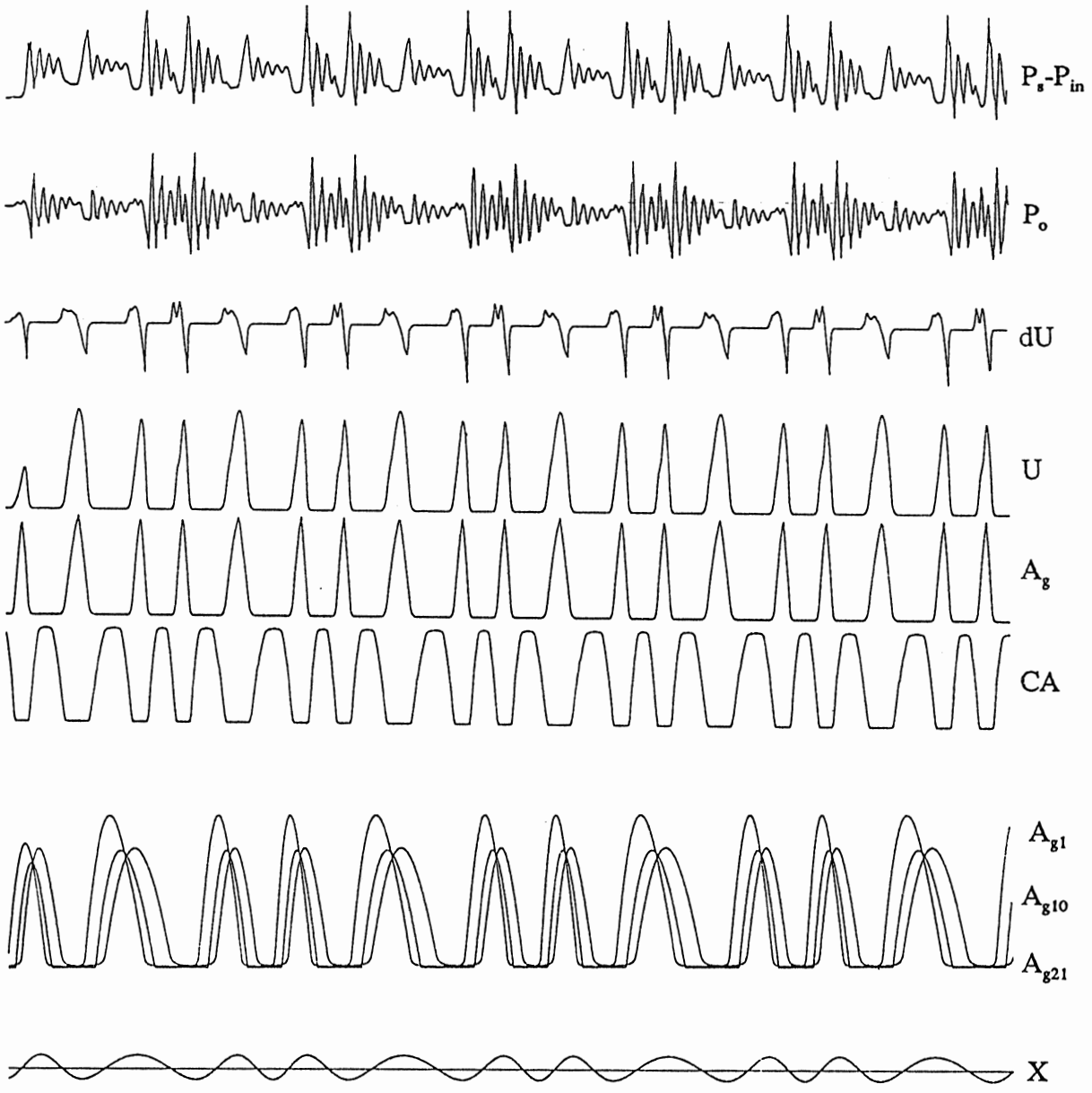


Figure 8. SPEAK generated subharmonic (1/3) array of signals. Assumes F_0 is constant for Equation 2 and varies for Equation 4, causing a constant phase delay between displacements, but uses 1/3 subharmonic to generate displacement slope changes.

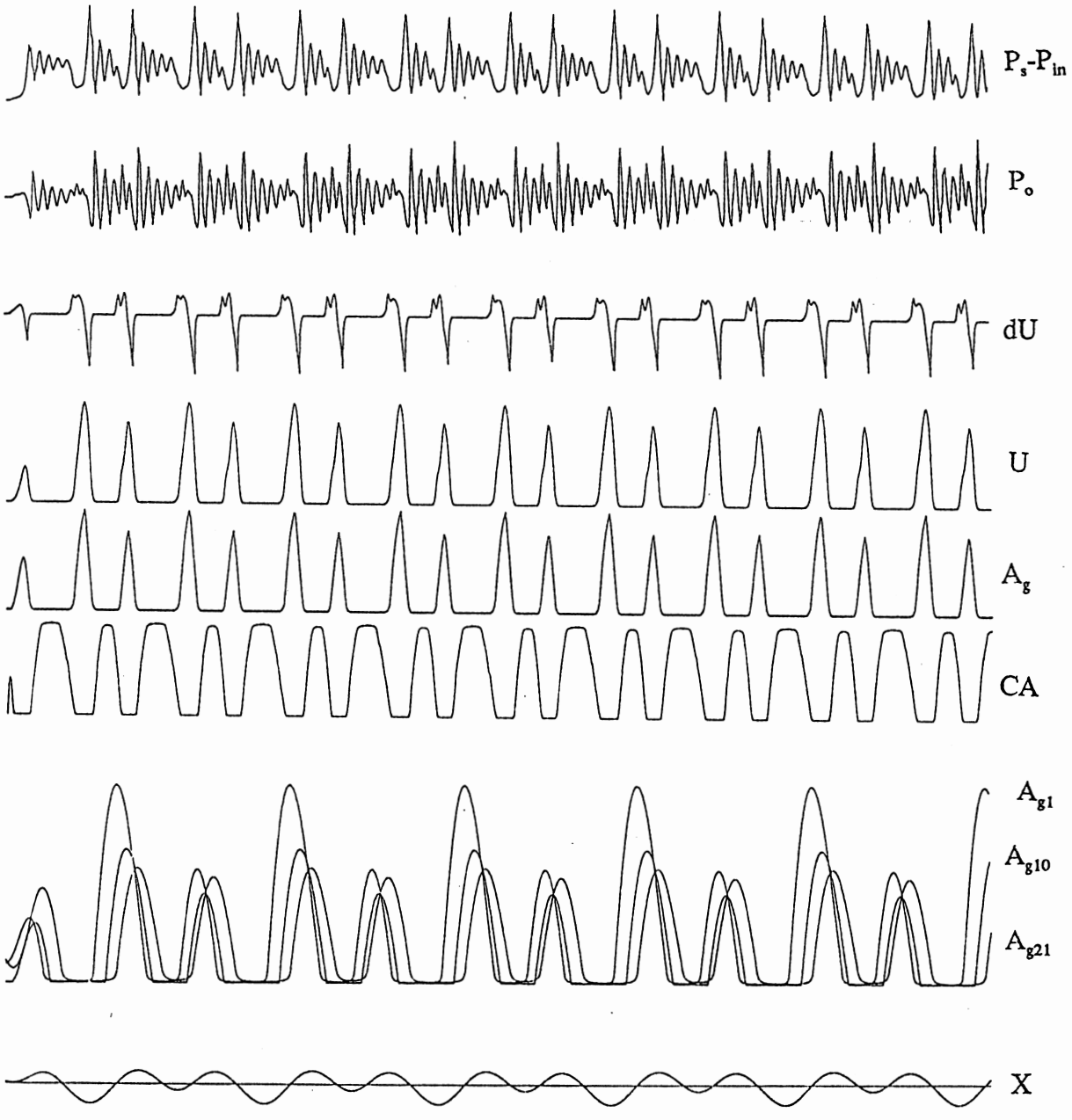


Figure 9. SPEAK generated subharmonic (1/2) array of signals. Assumes F_0' (time varying) is used for Equations 2 and 4, causing a constant phase delay between displacements, and amplitude modulation of the displacement.

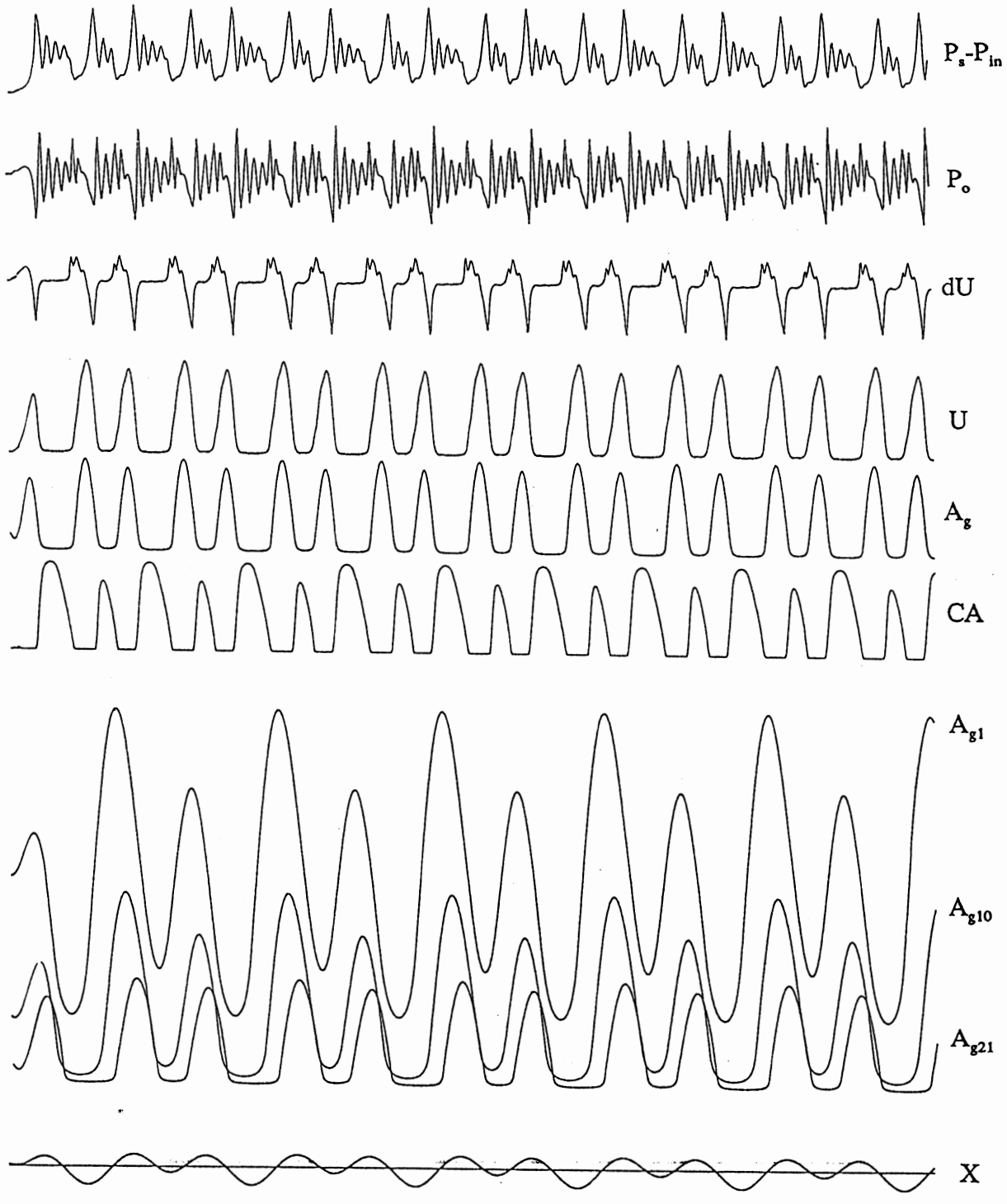


Figure 10. SPEAK generated subharmonic (1/2) array of signals. Assumes F_0' (time varying) is used for Equations 2, and 4, causing a constant phase delay between displacements, and amplitude modulation of the displacement. Lower folds do not close.

A Guide to Selecting A/D Hardware

Martin Milder, University of Iowa

Consider the data acquisition chain:

Source --> Amplification --> Filter(s) --> A/D Converter

Source

The source may be, e.g. Microphone, Electroglottogram (EGG), Photoglottogram (PGG), Glottal Flow, Pressure.

Amplification (optional)

Must meet requirements of converter (some converters have built-in gain stages).

Filter(recommended)

Low pass at 1/2 of the sampling rate (or less). Only if there is no high frequency energy can you do this. Some converters will low-pass automatically once the sampling rate is selected.

A/D converter

Specifications

A. Miscellaneous
- number of Channels.
- max sampling length.
- input range

B. Bandwidth
- highest sampling rate
- lowest sampling rate
- fixed or variable sampling rates.

C. Signal/noise
- total noise is the sum of all parts in the chain.
- more bits, the better.

Source Types

(1) Microphone

Purpose	Freq response	S/N (dB)
Auditory	20Hz - 20Khz	80
F0 Extraction	F0/4 - 8*F0	-
RMS	-	high
Jitter	F0/4 - 20*F0	-
Shimmer	F0/4 - 20*F0	high
Harm/Noise	F0/4 - 20*F0	high
Inverse Filter	3Hz - 8*F0	-

(2) EGG or PGG

Purpose	Freq response	S/N(dB)
F0 Extraction	F0/4 - 8*F0	-
Wave Shape	3Hz - 20*f0	high

A Guide to Selecting A/D Hardware

Martin Milder, University of Iowa

Consider the data acquisition chain:

Source --> Amplification --> Filter(s) --> A/D Converter

Source

The source may be, e.g. Microphone, Electroglottogram (EGG), Photoglottogram (PGG), Glottal Flow, Pressure.

Amplification (optional)

Must meet requirements of converter (some converters have built-in gain stages).

Filter(recommended)

Low pass at 1/2 of the sampling rate (or less). Only if there is no high frequency energy can you do this. Some converters will low-pass automatically once the sampling rate is selected.

A/D converter

Specifications

A. Miscellaneous
- number of Channels.
- max sampling length.
- input range

B. Bandwidth
- highest sampling rate
- lowest sampling rate
- fixed or variable sampling rates.

C. Signal/noise
- total noise is the sum of all parts in the chain.
- more bits, the better.

Source Types

(1) Microphone

Purpose	Freq response	S/N (dB)
Auditory	20Hz - 20Khz	80
F0 Extraction	F0/4 - 8*F0	-
RMS	-	high
Jitter	F0/4 - 20*F0	-
Shimmer	F0/4 - 20*F0	high
Harm/Noise	F0/4 - 20*F0	high
Inverse Filter	3Hz - 8*F0	-

(2) EGG or PGG

Purpose	Freq response	S/N(dB)
F0 Extraction	F0/4 - 8*F0	-
Wave Shape	3Hz - 20*f0	high

CBatch: A Software Program for Format-Independent Analysis of Acoustic Waveform Data

Paul H. Milenkovic
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
1415 Johnson Drive
Madison, Wisconsin 53706

Abstract

CBatch is a software program for the DOS operating system that can feed a sequence of acoustic data files through a user-supplied acoustic analysis program. CBatch takes care of all of the translating required to read and write a variety of widely-used waveform file formats. The analysis program is a software filter that reads and writes to the DOS standard input and output data streams, and CBatch intercepts these streams and redirects them to waveform files. The analysis program obtains information typically stored in file headers (sample rate, number of samples, data range and units) by writing keyword strings to standard output and by reading alphanumeric responses from standard input. The analysis program writes a keyword string to tell CBatch that it is read for waveform data, and then it reads blocks of waveform samples from standard input and it optionally writes processed blocks of samples (such as a pitch trace) to standard output.

It would be extremely useful to make acoustic analysis software portable across different speech software packages and different types of computer. Acoustic analyses include pitch, formants, glottal waveshape, long-term average spectrum, voice perturbation, word segmentation, the resynthesis of speech from parameters, and many other signal transformations of interest to the voice and speech community. Such portability would facilitate the dissemination of research results, the repeatability of research studies, as well as the development of standards.

There are two ways to establish a standard for making acoustic analysis software portable. One way is to establish a standard file format that acoustic analysis

programs could read and write. The other way is to develop a standard interface between the acoustic analysis program and a shell program that could translate the many existing file formats. This report concentrates on this second approach.

The concept is to implement the acoustic analysis algorithm as a software filter. The software filter has an exceedingly simple interface: it reads blocks of bytes from the operating system-designated standard input stream and it writes blocks of bytes to the standard output stream. The software filter has control over how many bytes it wants to read or write. A shell program intercepts the standard input and output data streams and takes care of what files are to be processed and in which format.

There is an additional embellishment to the software filter. The filter program makes requests of the shell program by writing alphanumeric keywords to standard output and it reads the shell's alphanumeric responses from standard input. This communication is primarily for obtaining information from the waveform file header – sample rate, data units, number of samples – without having to know the details of the header format. This communication can also be used for signalling. For example, the filter can tell the shell when it wants to switch over to reading and writing blocks of waveform samples.

In the usual situation of a main program calling a subroutine, the main program asks the questions and the subroutine provides the responses. In this situation, it is the suboordinate filter that is asking the questions and the supervisory shell that is providing the responses. This makes the filter much simpler because it only needs to ask for the things it needs to know in the order it needs to know them. It also makes the filter interface extensible. Adding new keywords does not invalidate existing filters which only know about the old keywords. Finally, it makes the interface interactive. The filter can indicate whether it wants both the input and output data streams attached to waveforms or whether it wants input only so it can write the output in its own format.

The CBatch program is a shell that employs the proposed filter interface to implement batch processing of waveform files in a variety of formats. CBatch is a program written in Turbo Pascal (Version 4 or later) that runs under DOS (Version 3 or later). The complete source will be made available through E-mail, an FTP server, or on a floppy disk: the source is ready but details of distribution are being arranged.

Which file formats does CBatch support?

CBatch inputs waveforms from

SPHERE is the format developed by National Institute of Standards Technology (NIST) for dissemination of the Texas Instruments-MIT-NIST (TIMIT) speech recognition data base on CD-ROM. It has an ASCII keyword free-format header.

CMU is the format employed in the beta-test (preliminary) release of the TIMIT CD-ROM. It has a fixed-format binary header. CMU refers to Carnegie-Mellon University, and the file format was developed in the course of speech recognition work at that institution.

NCVS92 is the format developed at the National Center for Voice and Speech (NCVS) for the dissemination of test data for acoustic analysis algorithms. It also has an ASCII keyword free-format header. This type of header is quite compatible with SPHERE, and there was discussion at the February 1994 Denver meeting that SPHERE should replace this format. This format contains keywords for data units and data range (translation from binary into real-world units like Volts or dB) which are desirable to add to SPHERE.

RIFF is the Microsoft multimedia waveform file format. The files typically wave the .WAV extension. Software packages sold with the popular sound cards for multimedia (Sound Blaster 16, Pro-Audio, Ensoniqs, etc.) employ this format.

Kay .NSP is the file format developed by Kay Elemetrics, identified by files with the .NSP extension. The Kay Linquistics Database CD-ROM as well as the planned Kay Voice Database are in this format.

UW XRMB is the file format used to collect both acoustic and pellet track signals in the University of Wisconsin X-ray Microbeam system. The files have the .DF extension.

SpeechStation is the Ariel/Sensimetrics software package for acoustic analysis on the IBM PC.

CSRE40 is the file format employed by Version 4 of the CSRE software package.

CSpeech is the CSpeech software package.

ILS is Interactive Language Systems, a software package widely used on PDP-11 and VAX systems.

CBatch automatically detects waveform files in the above formats. If autodetection fails (many packages uses ILS files without all the fields filled) or if a format is not on the list, header parameters may be entered on the CBatch command line.

CBatch writes to these formats (a command line parameter selects the format)

SPHERE format is readable by the Entropics Waves+ software packages plus any other package that reads TIMIT data.

NCVS92 is similar in concept to SPHERE, but contains additional keyword definitions for data units and range.

RIFF is compatible with a wide variety of software supplied with the popular PC sound cards.

CSpeech is included, not because I want to give a special status to CSpeech, but because I am the only source of information of how to get files into that format.

ASCII is alphanumeric text, desirable for the output of pitch traces for statistical analysis.

nohdr is raw binary without any header.

How do I invoke CBatch?

Invoke CBatch from the DOS prompt by entering

`cbatch` *params fname cmd*

where *params* is a list of parameters separated by spaces, *fname* is either a valid DOS path name (directory and file name) or wildcard specification (such as `*.wav`), and *cmd* is the name of your filter program followed by the parameters required by that program.

The entry *fname* specifies one or more waveform files. If the filter uses both input and output, those are the input files, and the output will go to files of the same name but in the current directory. That way you can analyze files on CD-ROM (such as TIMIT) and have the results go to files in the current directory on your fixed disk. If the filter specifies a file extension, the output will go to files having that extension.

The parameters are

`/H:`*h* overrides the autodetection of the input file header and strips off *h* 16 bit words from the beginning of the file.

`/O:`*o* overrides the input file header specification for conversion between offset binary and two's-complement numeric formats. For 12 bit offset binary, specify `/O:2048`.

`/R:`*r* specifies the number of bits of data resolution. If the input file header does not specify resolution, the value *r* is assumed. If the input file header specifies resolution, the data will be shifted to *r* bits.

`/B` overrides the input file header to specify that input data needs to be byte-swapped (for waveforms collected on SUN or Macintosh computers).

`/N:`*n* overrides the input file header and specifies that the data contains *n* channels of sample interleaved waveforms.

`/C:`*c* selects channel *c* from multiple channels where *c* starts at 1.

`/D:`*d* is the decimation factor. The parameter `/D:2` means that the input file was downsampled (decimated) to contain only every second sample, so the input waveform will be upsampled by repeating every sample.

`/s:`*s* overrides the input file header specification of the sampling rate in kHz.

`/r:`*r* overrides the input file header specification for the data range. For example, `/R:12` (12 bit resolution) and `/r:20` (data range of 20) means a sample value of -2048 corresponds to -10 (Volts or whatever other units are specified) and 2048 corresponds to 10.

`/u:`*uname* overrides the input file header specification of the units name (such as `/u:Volts` or `/u:ml/s`).

`/F:`*rname* means that the *fname* entry specifies one or more directories, and that *rname* designates the same file name that occurs in those directories. For example, if the files are `c:\rec001\speech.wav`, `c:\rec002\speech.wav`, ..., specify the parameter `/F:speech.wav` and use `c:\rec???` for your *fname* entry.

`/A:`*a* specifies the number of decimal digits accuracy for ASCII output.

`/T:`*t* specifies the time step for ASCII output in ms.

`/W:`*w* specifies an analysis window length in ms for ASCII output. The first output sample will be at $w/2$ ms. The purpose of this parameter is to align the ASCII output times with other analyses and does not otherwise control the filter doing the actual analysis.

`/R` specifies run-length ASCII output. A new ASCII value is output when the signal changes, but no more than every `/T:`*t* ms if the signal changes too frequently.

/outformat selects the output waveform format. Choices are **/nohdr**, **/NCVS**, **/SPHERE**, **/RIFF**, **/CSpeech** (default that can be changed by recompiling **cbatch.pas**), and **/ASCII**. The different file formats are explained above.

Example:

```
cbatch /ASCII /T:2 /R c:\speech\*.wav cptacf cptacf.sav
```

applies the pitch analysis filter **cptacf** to all the files in the directory **c:\speech** having the **.wav** extension. The pitch analysis takes its parameter settings from the file **cptacf.sav**. The analysis results will be in ASCII format, sampled in run-length fashion (when the value changes) but no more than once every 2 ms. The output files will have the same name as the waveform files, but they will go into the current directory and have the **.f0** extension.

How do I write a filter program?

The source code distribution includes the sample filter program **rectify.pas**, which performs a full-wave rectification of the input signal. It also includes the Turbo Pascal unit file **stdio.pas**, which simplifies use of the filter interface.

The basic structure of a filter program is to write keyword queries to standard output and to read alphanumeric responses to the queries from standard input. When this communication phase is done, write the query **initiate_binary_IO** (the procedure call **StdCmd('')** substitutes **initiate_binary_IO** for the null string) to initiate the data phase. Then simply read blocks of 16 bit samples from standard input and optionally write blocks of 16 bit samples to standard output: the number of samples to read and write is under your control. If you are doing only input, keep reading samples until standard input returns zero samples (this action resets the interface). If you are doing input and output, write zero samples to standard output as your last call (this also resets the interface).

CBatch recognizes these keywords. The first set of keywords are the communication phase dialog preamble: they must precede the other keywords and if used, they must be in the order listed below.

extension=*ext* specifies that the output file has extension *ext* (up to three letters).

input requests waveform data on standard input.

output requests waveform data on standard output.

The following keywords make up the body of the dialog and may occur in any order after the preamble.

`NSAMP?` returns the number of samples in the input file.

`RES?` returns the number of significant bits in the 16 bit waveform samples (controlled by CBatch `/R:`*r* parameter).

`sample_ms?` returns the interval between samples in ms.

`start_ms?` returns the time of the starting sample in ms (usually zero in batch mode – may have nonzero value when filter is invoked from the CSpeech program).

`stop_ms?` returns the time of the final sample in ms.

`units_name?` returns the input file units name (such as `Volts`).

`data_scale?` returns a value: multiply sample integer by this value to get real-world values in `units_name` units. This value is negative if the waveforms are inverted (for some types of microphone).

`fname?` returns the data file name, with the designated extension (from `extension=`*ext*) tacked on. If you have selected filter input only and are outputting in your own file format, use this query to get a file name. This is also useful if you want to do your own input formatting and have selected filter output only.

`sname?` returns the actual input file name without changing its extension. This is useful if you need to label analysis results with their file of origin.

`DOUBLE:`*textquery* prompts the user with the text string *textquery* for a decimal (double precision data type) value. This is useful if the filter needs to query the user for analysis settings. An alternative is for the filter to read such settings from the command line or from a file.

`sample_ms=`*s* sets the waveform sample interval in ms. Use this if you have selected output only and doing your own file read.

`buffer_ms=` sets the output file duration in ms. Use this if you have selected output only and are doing your own file read.

`units_name=`*uname* sets the output file units name. If you are doing pitch analysis, you may want to set `units_name=Hz`.

`range=`*r* sets the output file range (use `range = 1000` for pitch analysis).

Following the dialog body, the keyword `initiate_binary_IO` is transmitted by the procedure call `StdCmd('')`. For the use of `StdCmd` to transmit keywords and for the calls to decode responses, see the sample program `rectify.pas` and the unit `stdio.pas`.

How is the source to CBatch organized?

CBatch is a Turbo Pascal (Version 4 or later) program divided into a main program a set of Turbo Pascal units (separately compilable modules containing procedures and data type declarations). These modules are

`cbatch.pas` is the CBatch main program.

`strblock.pas` contains useful routines for manipulating character strings and processing file names. Some of these routines have equivalents in the libraries of later versions of Turbo Pascal, but including these routines gives downward compability with Version 4.

`wavein.pas` contains routines to input all the supported waveform file formats. You may want to examine this unit if you want to read waveform files directly without the filter interface.

`waveout.pas` contains routines to output a subset of the supported file formats. It is most important to be able to read as many formats as possible to access archival data. Support for output to more formats may be added.

`wavehdr.pas` contains header description and header validation procedures for the supported file formats. This unit is useful as a reference to the different file header.s

Special Report for the NCVS
and
The Denver Center for the Performing Arts

AN ASSESSMENT OF THE VIABILITY OF MULTIMEDIA FILE FORMATS FOR VOICE DATA USE

Timothy W. Curran, P.E., CEng, MIEI

February 1994

Introduction

In a broad sense, the first voice data file was a multimedia file, and so is any voice data file, regardless of format. This report will consider only those multimedia files known a "Red Book" audio files found on CD-ROM, and that type of RIFF file known as a WAVE file which contains PCM audio. These file formats are evaluated for potential use in voice research, speech and language pathology and vocal pedagogy. In the interest of brevity, discussion has been limited.

The adoption of multimedia file formats implies that multimedia hardware will also be adopted. In some cases this may be true. The acceptance of multimedia holds the promise of widespread availability of voice data files over computer networks and in computer databases, as well as widespread ability to convert the files into audio, all with equipment that can be purchased on a consumer budget.

Benson, Sage and Cook (1) have proposed their "triple-gateway methodology" for evaluating emerging technologies. Their evaluation is based on identifying elements of risk due to uncertainty. Since their methodology is new, a description is in order. In their words:

> "It is based on the proposition that a technology, to reach a mature stage in which it yields useful products or services, must pass through three gateways: the technology gateway, the systems-management gateway, and the market gateway.
>
> "Passing through the technology gateway requires research ability, innovation, technical merit, and a technical champion.

> "The management gateway includes technology management, finance, enterprise management, and standards. The market gateway, sometimes referred to as a "demand-pull" gateway, includes societal and consumer needs and receptiveness and general economic conditions.
>
> "Normally, a technology passes through the first stages of technology development before confronting the management and market gateways, so the reader might wonder why we treat the market and management gateways first. We do so to emphasize that most successful technologies are, in the final analysis, deployed because of "market-pull" rather than "technology push." But in whatever order the gateways are listed, the key point is that the technology must pass through all three gateways before it can be successfully deployed."

The following analysis duplicates the paragraphing and headings of the original journal article:

A. Market Gateway.

1.) New uses. This is a new use of an existing technology associated with the personal computer. This association may be positive or negative. The positive is illustrated by the new high performance RISC workstations that can be purchased with 16 bit multimedia audio hardware, and the accompanying software that uses standard PC multimedia audio files. The negative is illustrated by cheap computer game programs.

2.) User skepticism about improved performance characteristics. Since the current voice data file formats have existed for some time, some skepticism about the wisdom of adopting a new file format is easily understood. Because of their complex structure, multimedia file formats initially create more work for programmers.

For distribution on CD-ROM, however, "Red Book" audio (the standard of music CDs) is the preferred format for audio files. CD-ROM readers contain the circuitry to play the files, thus eliminating the need for a separate piece of digital-to-analog equipment. Kay Elemetrics publishes a phoneme database on CD-ROM with Red Book audio files.

3.) Requirement for behavior adjustment by user. Computers and analysis machines will insulate their users from any changes. The requirement for behavior change, however, will fall most heavily on the part of programmers and researchers who do their own programming. To fully utilize multimedia files, programmers must convert to thinking in terms of data objects, and their freedom to improvise must be replaced by the discipline inherent in multimedia standards.

4.) Competitive technologies. Existing file formats meet existing needs. Future needs, however, appear to be best met by multimedia files. A large number of word processing, telecommunications and database programs are compatible with multimedia file formats. In order to take advantage of this compatibility, existing file formats must be encapsulated in a shell that mimics some of the functions of multimedia files. It may be easier to convert the original file into a standard multimedia file.

5.) Unpredictable technological developments. The fall of computer prices, the wide acceptance of multimedia, and the advent of CPU and DSP chips of unprecedented power

may herald the inclusion of hardware suitable for use as voice analysis systems as a standard component of computer motherboards.

According to Benson, "... information is an increasingly important driving factor in our economy..." The ability of multimedia file formats to adapt to modern demands for information may be the key to their ultimate acceptance in the voice community. The hunger of voice teachers for information will lead to their acceptance of whatever technology they can afford. At the present, only multimedia can provide audio hardware for less than $200, and software that converts 386 and 486 PCs into spectrum analyzers is available at no cost on CompuServe.

According to Communications Week, June 28, 1993, the Multimedia Community of Interest is dedicated to promoting "standards that make it easier for users to transmit multimedia information over networks." It is composed of representatives from IBM, British Telecommunications plc, Deutsche Bundespost Telekom, France Telecom, Intel Corporation, Northern Telecom Ltd. and Telastra of Australia. Along similar lines, the MIME-1 standard makes it easier to transmit multimedia files over the Internet.

For educational use, both audio and laryngoscopic video can be combined on the same CD-ROM. Two companies known to the present author are selling software that allows CD-ROMS to be launched (used) by both PCs and Macintosh Computers.

These factors tend to reduce the uncertainty of the adoption of multimedia file formats by new users. Users of present file formats may not desire additional capabilities.

6.) Legal barriers. Not applicable.

B. Management Gateway

The voice community is composed of a large number of individuals in many career fields, and some professional, academic and business organizations. That portion of the voice community already involved in voice analysis may not desire a radical change in file format. Another portion, mostly voice teachers, voice coaches and singers, is not now active in voice analysis, but is interested. This group will have the least resistance to multimedia file formats.

An important consideration is database management. In order to automate the search and retrieval of voice data samples, the voice data files must be classified. If a new file is added to a collection, an automated procedure must be available to add attribution, annotation, classification and other information to the database. Files that provide for storage of this type of information are more valuable than files that do not. Multimedia files already have the capability to store this information in their very structure. The present author is currently working to establish a committee to develop standards for the insertion of attribution, annotation, classification and other

information into multimedia files to enhance their use in medicine and other scientific fields.

C. Technology Gateway

1.) Innovativeness of technology. Multimedia technology was very innovative when first introduced, but has now gained considerable acceptance. This factor does not significantly contribute to uncertainty.

2.) Number of constituent technologies. Since multimedia hardware and equipment sales continue to increase, the large number of constituent technologies supporting multimedia is not contributing to uncertainty.

3.) Manufacturing difficulties. Not applicable.

4.) Institutional changes required to introduce the new technology. This is a factor of great uncertainty. If the voice community continues to conduct business as it has in the past few years, there is no great motivation to accept change. Present file formats are adequate.

When the voice community recognizes the need to transmit voice data files over computer networks, or archive large amounts of voice data, or classify a large number of voice samples and keep them ready for instant retrieval, then the possibility will exist for adoption of more flexible file formats.

Conclusions and Author's Assessment

Multimedia file formats do not offer enough readily apparent benefits to entice users to abandon entrenched file

formats. Equipment manufacturers, however, may consider optional compatibility with multimedia file formats as inexpensive insurance to cover future market demands. If the IEEE issues a standard for incorporating attribution, annotation, classification and other information into multimedia files, they have the potential for becoming a preferred medium for short term archival use.

Multimedia file formats may also find use for data interexchange between incompatible file formats. The small number of voice file users does not justify the large investment necessary to produce software that will convert files from any one format to any other format. Commercial word processing conversion software converts files to an intermediate form which is then converted into the desired format. It would be possible for each file format sponsor to produce software to convert their own files into a standard multimedia intermediate format. The resulting intermediate files would be compatible with database programs for automated search and retrieval.

For new users with consumer grade equipment, multimedia files are the only files widely supported with digital audio utilities.

REFERENCE

1 Benson B, Sage A, Cook G. Emerging Technology-Evaluation Methodology: With Application to Microelectromechanical Systems. IEEE Transactions on Engineering Management 1993; 42:114-123.

Journal of Speech and Hearing Research, Volume 36, 1177–1190, December 1993

Effect of Microphone Type and Placement on Voice Perturbation Measurements

Ingo R. Titze
Department of Speech Pathology and Audiology, and National Center for Voice and Speech The University of Iowa Iowa City
and
The Recording and Research Center The Denver Center for the Performing Arts Denver, CO

William S. Winholtz
The Recording and Research Center The Denver Center for the Performing Arts Denver, CO

This study was conducted to explore the effects of microphone type (dynamic vs. condenser) and pattern (omnidirectional vs. cardioid) on the extraction of voice perturbation measures for sustained phonation. Also of interest were the effects of distance and angle between the source and the microphone. Four professional-grade and two consumer-grade microphones were selected for analysis. Synthesized phonation with different amplitude and frequency modulations at fundamental frequencies of 100 Hz and 300 Hz were presented over a loudspeaker. Human phonation was also included to test the validity of loudspeaker presentations. Three microphone distances (4 cm, 30 cm, 1 m) and three angles (0°, 45°, 90°) were used for microphone placement. Among the professional grade microphones, the cardioid condenser type had the smallest effect on perturbation measures. In general, condenser types gave better results than dynamic types. Microphones with an unbalanced output did not perform as well as those with balanced outputs. Microphone sensitivity and distance had the largest effect on perturbation measures, making it difficult to resolve normal vocal jitter at anything but a few centimeters from the mouth. Angle had little effect for short distances, but a greater effect for longer distances. These conclusions are preliminary because the sampling of microphones, distances, and signal types was very coarse. The study serves only to chart the course for future work.

KEY WORDS: microphones, voice perturbation, jitter, shimmer, recording

Voice perturbation analysis continues to be of interest in studies of vocal fold vibration and assessment of voice disorders. It has the potential for quantifying small aperiodicities in a voiced speech signal, which may be helpful in understanding some of the mechanisms by which vocal fold vibration may be disturbed. Perturbation analysis appears to be limited, however, to those signals for which an essentially periodic process is only mildly altered by amplitude or frequency modulations. For gross qualitative changes in the signal (bifurcations), the technique may be of lesser value.

Perturbation analysis has gone through a refining process in recent years. It has been suspected that some of the lack of consistency of perturbation measures found across laboratories and across similar subject groups (Karnell, Scherer, & Fischer, 1991) may be related to technical flaws in recording and processing of signals. Doherty and Shipp (1988) showed that some analog tape recorders can greatly inflate vocal jitter (cycle-to-cycle irregularity in fundamental frequency) and vocal shimmer (cycle-to-cycle irregularity in amplitude). Digital audio recording is now commonly used, which essentially eliminates the contamination of perturbation measures by wow and flutter of the tape drive.

Other problems in perturbation analysis relate to the appropriate length of the analysis window (Karnell, 1992), the method of extraction used (Milenkovic, 1987; Titze & Liang, 1993), and the interactions between amplitude and frequency measures (Hillenbrand, 1987).

0022-4685/93/3606-1177$01.00/0

As in any high-fidelity audio recording, attention also needs to be paid to the microphone, the conditioning amplifiers and filters, and the recording environment. This article deals with two problems, microphone type and microphone placement relative to the source. Questions of interest are (a) whether the extraction of F_0 and amplitude perturbation is more accurate or consistent with a condenser microphone or a dynamic microphone, (b) if directionality of the microphone is an important factor, (c) how perturbation measures are affected by mouth-to-microphone distance, and (d) whether off-axis placement is an important factor.

Interactions between room environment, recording hardware, and extraction software can be extremely complex. For example, the interaction matrix (signal type × perturbation type × perturbation extent × microphone type × distance × angle × F_0 × room noise) can easily have several thousand items, even if only a few cases are selected for each variable. Therefore, to limit the scope of this first study, only a few typical angles and distances were used in combination with a few typical microphones. As a result, none of the questions posed above are answered unequivocally. It was felt, however, that a broad paintbrush approach was needed to determine where future efforts should be placed.

Method

Recording Environment

The experiment was conducted in an IAC isolation booth, 3.1 m deep by 3.5 m wide by 2.4 m high. The booth is sound treated, but not anechoic, so that source and microphone placement can affect the measurement. Both human speakers and a loudspeaker were used to generate the acoustic signal. The locations of the microphones relative to the source are shown in Figure 1. There were three angles (0°, 45°, and 90°) for each of three distances (4 cm, 30 cm and 1 m). A distance of 1 m was chosen as a compromise to approach asymptotic (far field) conditions on the one hand but maintain a reasonable signal-to-noise ratio on the other hand. The 30-cm distance was chosen as a mid-range value commonly used in acoustic voice recordings (Schutte & Seidner, 1983), and 4 cm was chosen for possible head-mounted microphone applications. Other distances, such as the common 15-cm distance, could have been added, but results for such intermediate distances can be found by interpolation of the present data.

Ambient sound pressure level (SPL) of the booth, measured with a B&K 2230 SPL meter on the linear weighting scale (20 Hz–20 kHz), was 53 dB near the center of the booth. This is typical for economically feasible laboratory environments.

To test the resonance effects of the booth, a frequency sweep of 20% around the fundamental frequency (F_0) of 300 Hz was conducted with a sinusoid at the three distances from the source (0° angle and an SPL of 80 dB). The frequency sweep corresponded to the maximum frequency modulation imposed in later experiments. Measurements indicated that for these modulations the intensity varied as follows: 0.6 dB at 4 cm, 2.8 dB at 30 cm, and 4.4 dB at 1 m. This is probably an underestimate of the room reverberation problem for nonsinusoidal voice signals, given that harmonics can carry a significant portion of the energy. A portion of this amplitude variation can perhaps be attributed to nonuniform loudspeaker response, but the sizable increase in variation with distance suggests that amplitude perturbations can result from source frequencies being selectively reinforced by room resonances. A similar effect was demonstrated with regard to the resonances of the vocal tract (House, 1960; Hillenbrand, 1987; Horii & Hata, 1988). Room resonances will vary with the size and type of room. As indicated by the measurements reported later, positioning the microphone nearer to the source will minimize the variations. A separate study of room effects on perturbation measures is needed to go beyond this simple rule of thumb.

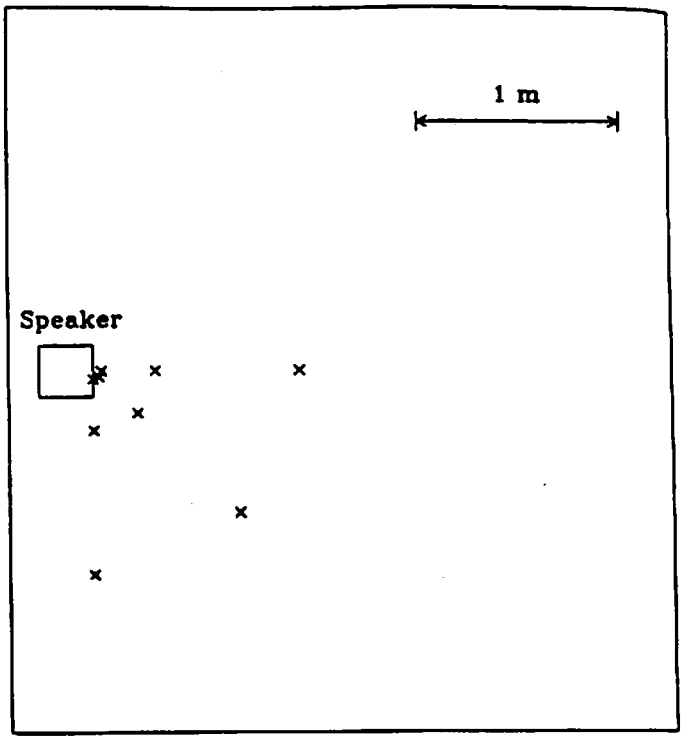


FIGURE 1. Diagram of loudspeaker and microphone placements in an IAC booth, drawn to scale (top view).

Microphone Type

Four professional-grade microphones were chosen. They all had electrically balanced outputs and a flat frequency response (< ±1.4 dB) over the ranges of frequencies used in this experiment. The microphones are identified as follows:

#1–AKG 451EB CK22 condenser omnidirectional
#2–AKG 451EB CK1 condenser cardioid
#3–EV DO54 dynamic omnidirectional
#4–AKG D224E dynamic cardioid

In addition, two consumer-grade microphones with electrically unbalanced outputs were chosen for limited analysis and comparison to the professional-grade microphones:

#5–Realistic 33-985 dynamic omnidirectional
#6–Realistic 33-1063 tie clip miniature condenser omnidirectional

TABLE 1. Sensitivity of microphones used in this study.

Microphone	Type	Sensitivity (dB)	Grade
1	condenser omnidirectional	−49.56	professional
2	condenser cardioid	−48.20	professional
3	dynamic omnidirectional	−68.69	professional
4	dynamic cardioid	−70.20	professional
5	dynamic omnidirectional	−75.24	consumer
6	condenser omnidirectional	−76.03	consumer

Sensitivity of the microphones was determined by placing them 4 cm from the loudspeaker at an 0° angle, which produced a SPL of 80 dB at 200 Hz. The output of the preamplifier minus the constant gain (60 dB) was recorded. Results are shown in Table 1. The two professional-grade condenser microphones (1 and 2) were clearly the most sensitive, and the consumer-grade condenser microphone (6) was the least sensitive. It will be seen later that microphone 2 gave consistently the best results, with microphone 1 being a close second.

Test Signals

A computer synthesis program (Titze & Liang, 1993) was used to generate pulsatile glottal flow waveforms having different types of modulation (see flow chart in Figure 2). On the top left of the diagram are controls for frequency modulation (FM), including the modulation index, modulation frequency, and carrier frequency (F_0). On the top right are similar inputs for amplitude modulation (AM). In the middle blocks, the instantaneous phase and the magnitude envelope of a glottal pulse are computed and noise is added. To

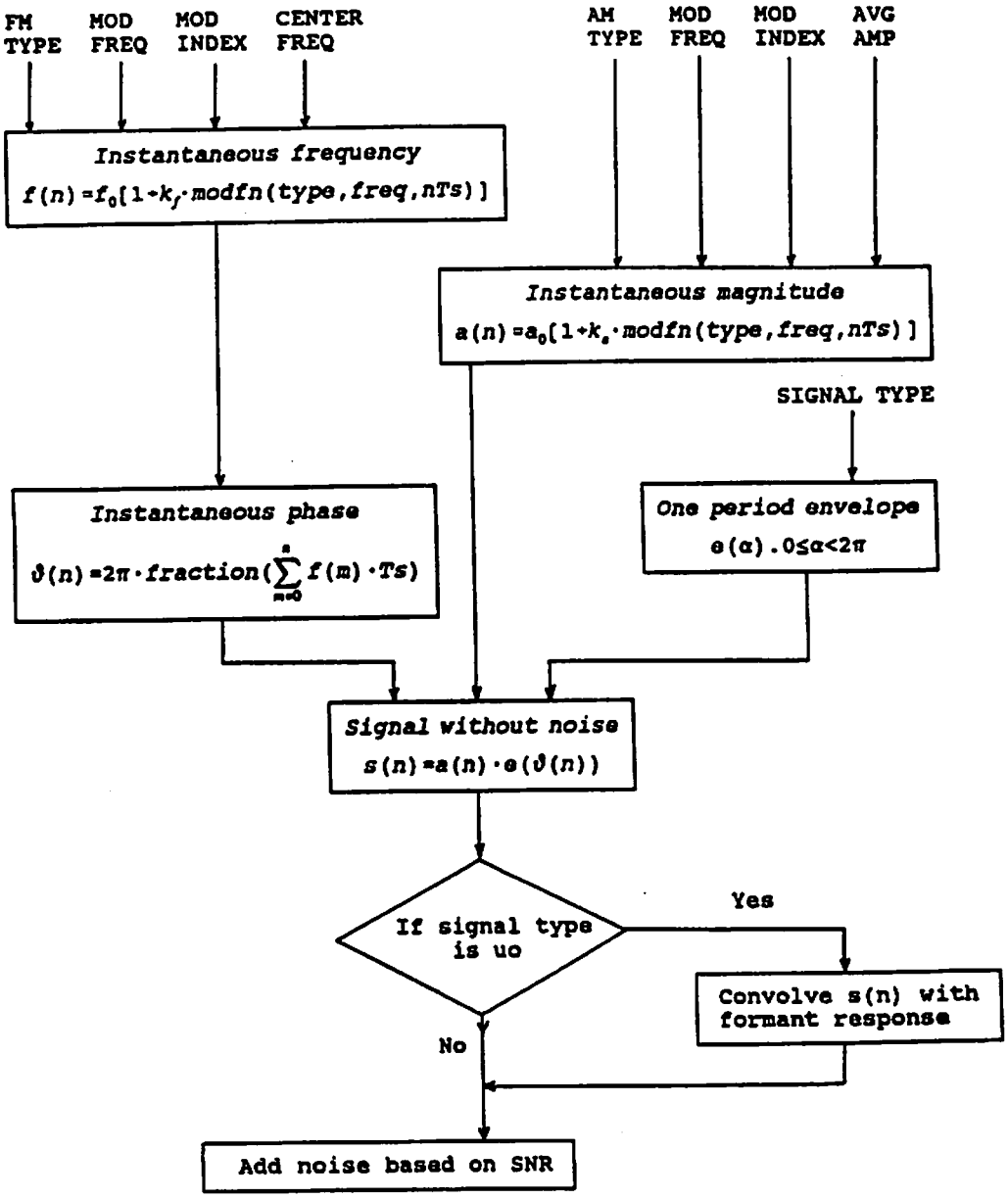


FIGURE 2. Flow chart of computer program to synthesize modulated speech signals.

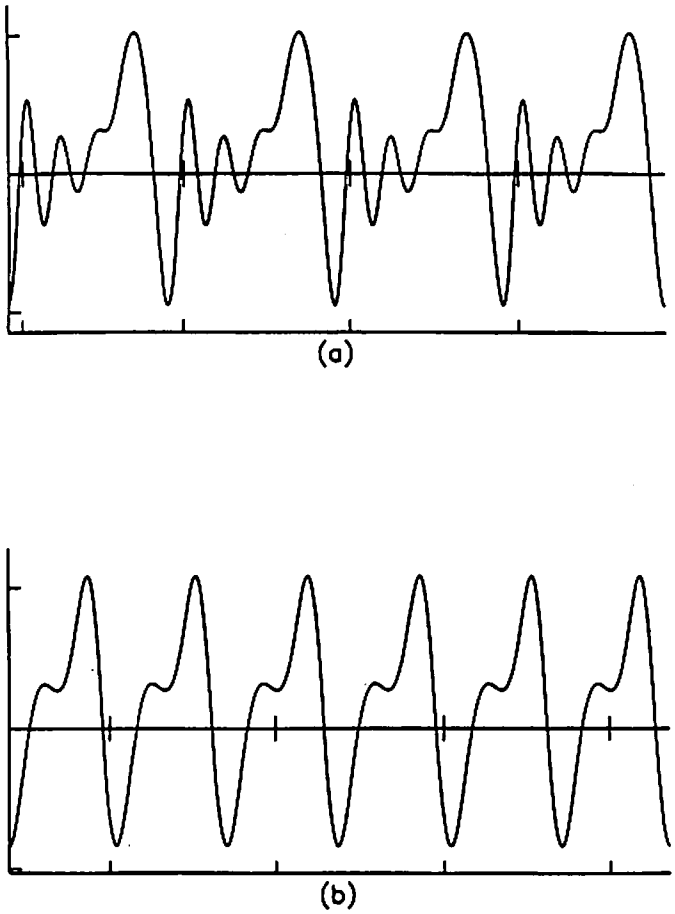


FIGURE 3. Stimulus signal with no imposed modulation. (a) F_0 at 100 Hz and (b) at F_0 at 300 Hz. The formant frequency is 470 Hz.

simulate jitter or shimmer, an $F_0/2$ (subharmonic) modulation was used to vary adjacent pulse period and amplitude, respectively.

To make the waveform somewhat speech-like, a single formant was simulated by convolving a glottal pulse with a filter response (see bottom of Figure 2). The formant frequency was set to 470 Hz, a number that has no integer relation to any F_0 used. This avoided any strong boost that a single harmonic might get from the formant filter. Additional formants could have been added to make the signal more speech-like, but we felt that this was a reasonable compromise between using a spectrally rich waveform and a spectrally poor waveform (which is seen in some falsetto and female productions).

Two groups of signals with identical modulations were generated. One group had an F_0 of 100 Hz and the other an F_0 of 300 Hz. Each signal was 6 seconds in length. Figure 3(a) shows the 100 Hz signal and Figure 3(b) shows the 300 Hz signal, both without imposed modulations.

In each F_0 group there were four types of modulation: (a) AM (amplitude modulation) at 10 Hz, (b) AM at $F_0/2$, the first subharmonic, (c) FM (frequency modulation) at 10 Hz, and (d) FM at $F_0/2$, the first subharmonic. All modulations were sinusoidal. The 10-Hz AM was synthesized at 50% and 5% modulation index to correspond to ranges observed for vocal tremor (Winholtz & Ramig, 1992) and normal voice, respectively. The 10-Hz FM had 20% and 2% modulation indices, realizing that FM is generally smaller than AM in speech. All 10-Hz modulations maintained at least a 10:1 ratio in carrier frequency to modulation frequency.

For the subharmonic modulations, however, the ratio was 2:1. This short-term AM and FM modulation had indices of 5% and 0.5% to simulate shimmer and jitter (Scherer, Gould, Titze, Meyers, & Sataloff, 1988). Random period fluctuations could have been used, but in our experience these are a bit more difficult to control synthetically. There are problems with discontinuities within and across cycles and there are further problems with knowing what the precise theoretical value is. The $F_0/2$ modulations give a regular stream of reversals in the F_0 or amplitude contour (every cycle), but they obviously do not capture every feature of jitter and shimmer.

Included in each F_0 group was one signal with no imposed modulation. This signal was used to determine the noise of the system and the errors associated with perturbation extraction.

For convenience in later presentation, the signals were transferred from the computer through a DSC-200 16-bit D/A converter (at 20-kHz sampling frequency) to a Panasonic SV-3700 DAT recorder. (Direct digital transfer was not possible because of present format incompatibility.)

Hardware for Signal Presentation

The DAT recorder was used as the signal source for a Technics SU-V303 power amplifier and a 4-inch loudspeaker (Auratone 5C). The sound pressure level (SPL) was adjusted to 80 dB at a distance of 4 cm from the loudspeaker with the 100-Hz signal (no modulation). It was not re-adjusted for the remainder of the experiment. The loudspeaker was chosen over human phonation (for most of the experiment) because it offered a wide range of control and was more consistent in presentation of the acoustic stimulus. For example, when a synthesized phonation signal with no perturbation was presented over the loudspeaker, it produced baseline perturbation measures an order of magnitude lower than average normal human phonation. This allowed for better assessment of the influence of system noise and for better verification of the imposed signal modulations. However, to test the differences between a human speaker and a loudspeaker as a source of perturbation, some comparable measurements on humans were made. These will be described later.

Hardware for Recording

The microphone signals were preamplified (ATI M-1000), high-pass filtered at 60 Hz (24 dB/oct linear phase) and sampled by a DSC-200 16-bit A/D converter (20-kHz sampling with a 10-kHz antialiasing filter). To utilize the full range of the A/D converter, the gain of the conditioning amplifier (DSC-240) was adjusted for "0" VU with the unmodulated signal each time a microphone was repositioned. After the initial setting, the gain was unchanged for the remainder of the experiment for that position. A 3-sec segment of the total 6-sec stimulus was digitized in each case, avoiding start-up and release transients.

TABLE 2. Direct analysis of the synthesized signals used for stimulus.

Imposed modulation		Amplitude measures (%)			
		F_0 = 100 Hz		F_0 = 300 Hz	
		CV	P1	CV	P1
None		0.00	0.00	0.08	0.12
Amplitude	50%	31.45	17.23	34.78	6.55
	5%	3.39	1.84	3.53	0.66
Frequency	20%	16.12	8.30	0.96	0.40
	2%	2.00	1.07	0.27	0.70
Shimmer	5%	1.74	3.47	2.43	4.83
	0.5%	0.18	0.35	0.26	0.49
Jitter	5%	0.52	1.01	0.85	1.67
	0.5%	0.07	0.13	0.10	0.17
		Frequency measures (%)			
		F_0 = 100 Hz		F_0 = 300 Hz	
		CV	P1	CV	P1
None		0.00	0.00	0.00	0.00
Amplitude	50%	0.09	0.08	0.02	0.01
	5%	0.00	0.00	0.00	0.00
Frequency	20%	12.05	11.06	13.70	2.59
	2%	1.34	0.75	1.41	0.26
Shimmer	5%	0.05	0.11	0.02	0.04
	0.5%	0.01	0.01	0.00	0.00
Jitter	5%	1.51	3.06	1.80	3.60
	0.5%	0.15	0.30	0.18	0.36

Note. Modulation frequencies were 10 Hz, except for jitter and shimmer, which was an $F_0/2$ subharmonic modulation.

Analysis

Software

A software package called GLIMPES (Glottal Imaging by Processing External Signals), was used to analyze a random 2-sec segment of the digitized signals, again avoiding the earliest and latest cycles. Since the stimulus used for presentation was highly repetitive, only the length of the token was critical for comparison in analysis. The F_0 extraction algorithms are of two types: (a) single-event detection (peaks or zero crossings) with interpolation, and (b) waveform matching between adjacent cycles with interpolation. The accuracy of extraction of these measures is being reported (Titze & Liang, 1993). Waveform matching, which determines the period on the basis of the least square error between adjacent-cycle waveforms, gives the best results for F_0 extraction in the presence of additive noise and amplitude modulation. Single-event detection, and in particular zero-crossing detection on the low-pass filtered signal, gives better results only when F_0 is modulated more than a few percent. We used the waveform matching technique in this experiment. The technique is described in detail by Milenkovic (1987) and by Titze and Liang (1993).

Amplitude and frequency perturbation measures were obtained. The measures included CV (coefficient of variation), defined as the zero-order RMS (root mean squared) perturbation according to the Pinto and Titze (1990) nomenclature, and P1 (perturbation one), defined as the first-order MR (mean rectified) perturbation:

$$CV = \frac{100}{\bar{x}}\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2} \quad (1)$$

$$P1 = \frac{100}{(N-1)\bar{x}}\sum_{i=2}^{N}|x_i - x_{i-1}| \quad (2)$$

In the above equations, the variable x can be either frequency or amplitude, i is the period index, and $\bar{x}$ is a mean value for normalization to percent.

Results

Direct analysis of the synthesized computer files (without playback, recording, and redigitization) provided the reference for comparison with the corresponding recorded microphone signals. Table 2 illustrates the results for this direct analysis. The top half of the table is for amplitude measures and the bottom half for frequency measures. For the signal with no imposed modulation (first line in each half), amplitude perturbation ranged between 0.00 and 0.12% over both values of F_0, and frequency perturbation was undetectable to 0.01% accuracy. This is essentially the software (extraction)

TABLE 3. Extracted perturbations from microphone signals for stimuli with no imposed modulation, 0° angle, and 4 cm from source.

	Amplitude measures (%)			
	F_0 = 100 Hz		F_0 = 300 Hz	
Microphone	CV	P1	CV	P1
1	0.19	0.22	0.15	0.20
2	0.10	0.11	0.13	0.17
3	0.34	0.34	0.25	0.34
4	0.34	0.34	0.27	0.33
5	1.53	1.67	1.28	1.69
6	0.98	1.00	0.73	0.93
	Frequency measures (%)			
	F_0 = 100 Hz		F_0 = 300 Hz	
	CV	P1	CV	P1
1	0.01	0.02	0.04	0.05
2	0.01	0.01	0.02	0.03
3	0.02	0.02	0.05	0.09
4	0.02	0.03	0.03	0.05
5	0.10	0.14	0.18	0.27
6	0.05	0.07	0.10	0.14

jitter and shimmer, which is low enough not to inflate the imposed modulations significantly.

There is no direct relation between modulation index and the CV and P1 measures. Modulation index is defined as

$$Mod = \frac{x_{max} - x_{min}}{x_{max} + x_{min}}, \quad (3)$$

where x_{max} and x_{min} are maximum and minimum values of either frequency or amplitude. This index is used for convenience in synthesis. It is easier, mathematically, to impose a given modulation than a given CV or P1. As long as both are calculated and measured, there is no loss of information. Note that for 50% AM (at the 10-Hz modulation frequency), the CV for amplitude is about 30%. The P1 varies from about 17% at an F_0 of 100 Hz to about 7% at an F_0 of 300 Hz (second line in Table 2). Variations in P1 result from the fact that adjacent cycle differences are minimized when the ratio of modulation frequency to F_0 decreases. Note that both CV and P1 scale downward appropriately by a factor of 10 when modulation is reduced by a factor of 10.

Cross-modulation between AM and FM is always a problem in speech signals (Hillenbrand, 1987). In the fourth line of the top half of the table, we see that an imposed 20% FM results in large amplitude perturbations, even though no AM was imposed. These cross-modulations result from different harmonics moving through formants at different fundamental frequencies. At 100 Hz, for example, the fifth harmonic is within 30 Hz of the 470-Hz formant. A 20% frequency variation (up and down) will sweep this harmonic through the formant, causing large amplitude fluctuations. At F_0 = 300 Hz, on the other hand, a 20% frequency variation will sweep neither the fundamental nor the second harmonic through the formant frequency. The amplitude perturbations are correspondingly smaller at 300 Hz in Table 2.

The cross-modulation is less severe when AM is imposed. Note that in the second line of the bottom half of the table, a 50% AM causes values of CV and P1 that are all less than 0.1%. Thus, although there is no large frequency-amplitude interaction that corresponds to the amplitude-frequency interaction of vocal tract formants, small amounts of cross-modulation do occur. They were also reported by Hillenbrand (1987) and seen to be dependent on the analysis and synthesis methods.

Having assessed the baseline accuracy of the extraction software and the possible hazards due to cross-modulation, our next step was to assess the interactions between hardware and software. In particular, the effect of the six microphones and their placement were of interest.

Microphone Type

Table 3 shows the perturbation measures when the entire electroacoustic link (DAC, power amplifier, loudspeaker, microphone, preamplifier filter, DAT recorder, and ADC) was included. No modulations were imposed. Hence, these are the baseline measures for the hardware-software combination. The effect of the electroacoustic link, and particularly each microphone, can be assessed by comparing these numbers to the two rows labeled "None" in Table 2. With professional grade microphones (1–4), the overall amplitude perturbations were on the order of 0.1%–0.3% and the frequency perturbations were an order of magnitude lower. Perturbations with the dynamic microphones (3 and 4) were generally higher than with condenser types (1 and 2). With consumer-grade microphones (5 and 6), baseline perturbations were about three times higher, on average, than with the professional grades. This may be caused by their electrically unbalanced outputs, which allows increased electrical noise to contaminate the signal, either by improper shielding

TABLE 4. Comparison of microphone-extracted modulation measures (CV for 100 Hz modulations and PF for $F_0/2$ modulations) for different types of imposed modulation, 0° angle and 4-cm distance from source.

	Imposed modulation							
	AM (5%, 10 Hz) CV		AM (5%, $F_0/2$) P1		FM (2%, 10 Hz) CV		FM (5%, $F_0/2$) P1	
F_0	100 Hz	300 Hz	100 Hz	300 Hz	100 Hz	300 Hz	100 Hz	300 Hz
Direct	3.39	3.53	3.47	4.83	1.34	1.42	3.03	3.60
Mic								
1	3.42	3.56	4.77	3.68	1.36	1.42	4.52	3.38
2	3.36	3.53	2.78	4.32	1.34	1.41	3.09	3.64
3	3.43	3.55	5.37	3.56	1.37	1.41	4.23	3.40
4	3.35	3.53	3.81	4.72	1.32	1.40	2.79	3.27
5	3.68	3.72	3.90	4.09	1.32	1.40	2.44	4.19
6	3.49	3.58	5.03	4.39	1.35	1.41	2.66	2.98

from electromagnetic interference (giving greater additive noise) or from improper isolation from ground loops (producing modulations of 60 Hz and multiples of 60 Hz). Much of the modulation was not eliminated by the 60-Hz HP filter because the attenuation was only 3 dB at 60 Hz. To reduce the effects of ground loops entirely with HP filtering, the cutoff frequency would have to be higher than 60 Hz, but this would then interfere significantly with the $F_0/2$ subharmonic at 100 Hz. Since the purpose of the HP filter was primarily to reduce low frequency acoustic noise in the IAC booth, a compromise of 60 Hz was struck.

Table 4 presents another comparison of extracted perturbations with and without the electroacoustic link, but this time some modulations are imposed. To shorten the table, CV values are given for the 10-Hz modulations and P1 values for the $F_0/2$ modulations. These are deemed the "most appropriate" measures for the type of modulation. For the 10-Hz amplitude and frequency modulations, the values extracted with professional grade microphones were all within ±2% of the values obtained from direct analysis. With the consumer grade, microphone 5 (the dynamic omnidirectional) had the effect of inflating the 10-Hz amplitude modulations slightly (first two columns), but the frequency modulations were not affected much (columns 5 and 6). Recall from Table 3 that this microphone had the largest baseline amplitude measures when no modulations were imposed.

For the $F_0/2$ modulations, there was greater variability. About half of the measurements were below the direct analysis values (columns 3 and 4 and columns 7 and 8). This may seem strange. One does not expect electronic equipment to *deflate* perturbation values. The problem is likely to be phase distortion (discussed below), which has a strong effect on $F_0/2$ modulations. Because of the complex nature of the waveform, modulation frequencies that approach the carrier frequency change the waveshape unevenly from cycle to cycle. If the exact waveshapes are not preserved by the microphones or the amplifiers, different event locations will be detected. The waveform matching technique may smooth out some of these variations of individual events. This may be one of the drawbacks of this method of extraction, but it has many other advantages (Titze & Liang, 1993). In general, the $F_0/2$ measures always appeared to be more variable across microphones than the measures for the low-frequency modulations.

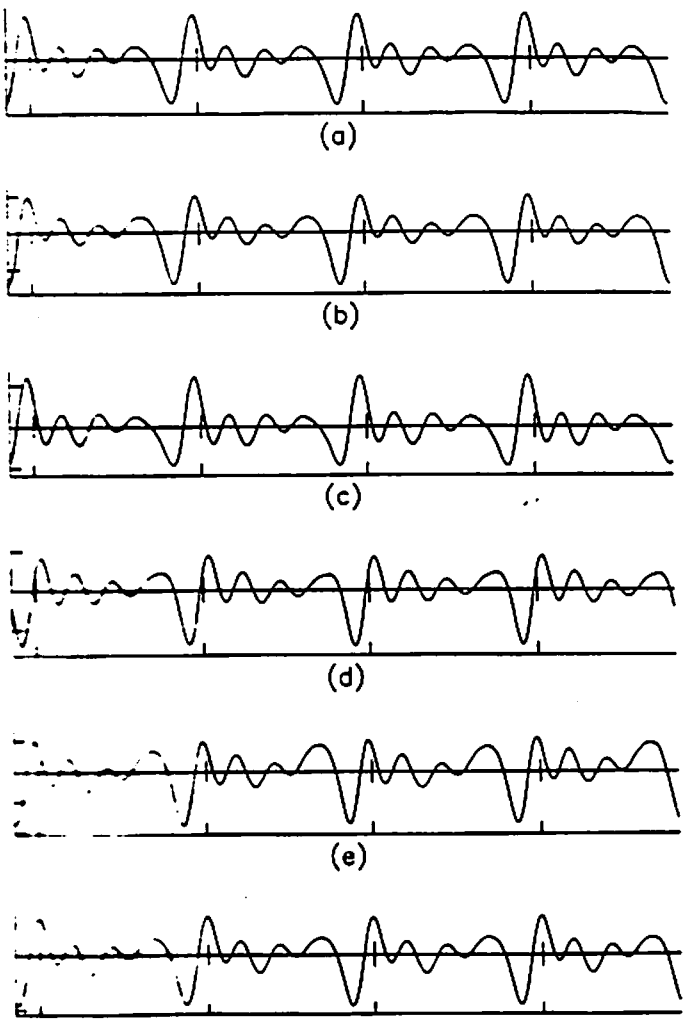


FIGURE 4. Recorded signals from microphones for stimulus in Figure 3(a) at 0° angle and 4 cm distance. Parts (a)–(f) correspond to microphones 1–6.

TABLE 5. Perturbation measures of six microphones from identical sections of phonation over 4 normal-speaking subjects (0° angle, 15 cm).

Subject	F_0	Mic	Amplitude		Frequency	
			CV	P1	CV	P1
M1	121 Hz	1	4.01	0.92	0.34	0.17
		2	4.36	1.16	0.34	0.17
		3	3.99	1.00	0.34	0.17
		4	3.85	0.82	0.35	0.17
		5	3.82	0.90	0.34	0.18
		6	4.46	1.00	0.34	0.17
M2	85 Hz	1	10.01	3.30	0.99	0.36
		2	9.97	3.29	0.99	0.33
		3	9.19	2.55	0.98	0.59
		4	7.60	2.19	1.00	0.34
		5	6.21	1.74	0.98	0.35
		6	9.92	2.98	0.99	0.34
F1	246 Hz	1	5.82	2.66	0.54	0.59
		2	5.68	3.43	0.54	0.66
		3	5.75	2.41	0.54	0.95
		4	7.13	3.24	0.54	0.71
		5	5.58	2.66	0.54	0.64
		6	5.66	3.22	0.54	0.60
F2	164 Hz	1	2.12	0.76	0.41	0.27
		2	2.12	0.84	0.41	0.27
		3	2.04	0.84	0.41	0.27
		4	2.27	0.79	0.42	0.28
		5	2.18	0.95	0.41	0.27
		6	2.20	0.74	0.41	0.26

Phase distortion varied between microphone types as demonstrated in Figure 4. The signals were recorded for F_0 = 100 Hz, 0° angle, and 4 cm distance. The input signal to the loudspeaker, for comparison, is Figure 3(a). All of the microphones distorted the signals differently. For example, compare waveforms (a) and (d), representing the omnidirectional condenser versus the cardioid dynamic microphones, respectively. Not only is there a significant phase delay in (d), as seen by the line-up at time = 0, but the largest biphasic pulse is less symmetric about the zero axis. The major downward peak is greater than the following upward peak. This difference between the major upward and downward peaks is even greater when waveforms (c) and (d) are compared. Similar distortions occurred also at 300 Hz.

Unfortunately, all of the recorded signals show evidence of general loss of frequency response with respect to the input voltage to the loudspeaker. The best example of this is the suppressed large positive peak *preceding* the major negative peak (compare any waveshape in Figure 4 with Figure 3a). This is the portion of the waveform that represents the glottal flow pulse (modified by the formant). Loss of low-frequency response of the loudspeaker, both in magnitude and phase, degrades the signal before it is recorded.

To verify that the phase distortion was not unique to loudspeaker presentation, however, experiments with human speakers were also conducted. Because human subjects could not repeat identical utterances, multi-channel simultaneous recordings were made onto an eight-channel DAT recorder prior to computer analysis. One experiment was designed to determine any near-field effects on sustained phonation. Three microphones of the same type were positioned at three angles (0°, 45°, 90°) at a distance of 4 cm from 1 subject. The experiment was then repeated at 15 cm. Visual inspection of the waveform shapes and perturbation analysis revealed no obvious microphone differences for these positions, at least not for the vowel [a] used throughout the experiments.

In another experiment, the six microphones were positioned in a cluster at a 15-cm distance and 0° from the mouth of 1 speaker. The greater distance was needed to accommodate the cluster, which measured about 6 cm in diameter. To determine if there were any spatial effects within the cluster, two of the microphones were then reversed in position (about 6-cm separation). A comparison of the resulting waveshape indicated no significant visual differences, and perturbation measures were also within baseline extraction errors.

To quantify the perturbation measures with the microphone cluster for several speakers, 4 normal-speaking subjects (2 male, 2 female) produced sustained phonation of an [a] vowel (again at 0° angle, 15 cm from the subject's mouth). Table 5 presents the results of perturbation analysis on the identical segments of phonation. Note that some large outliers are found among the measures across microphones. These occur most often with microphones 3, 4, and 5, the dynamic types. Obvious examples are the amplitude CV's for subjects M2 and F1 for microphone 4. Recall that this was the microphone with the largest phase delay (part d in Figure 4).

Figure 5 shows some recorded waveshapes for subject M1, a male phonating at 121 Hz. It can be seen that the waveform distortions are similar to those found for the 100-Hz loudspeaker presentations. In particular, note the phase delay with microphone 4 (part d of the figure). Also

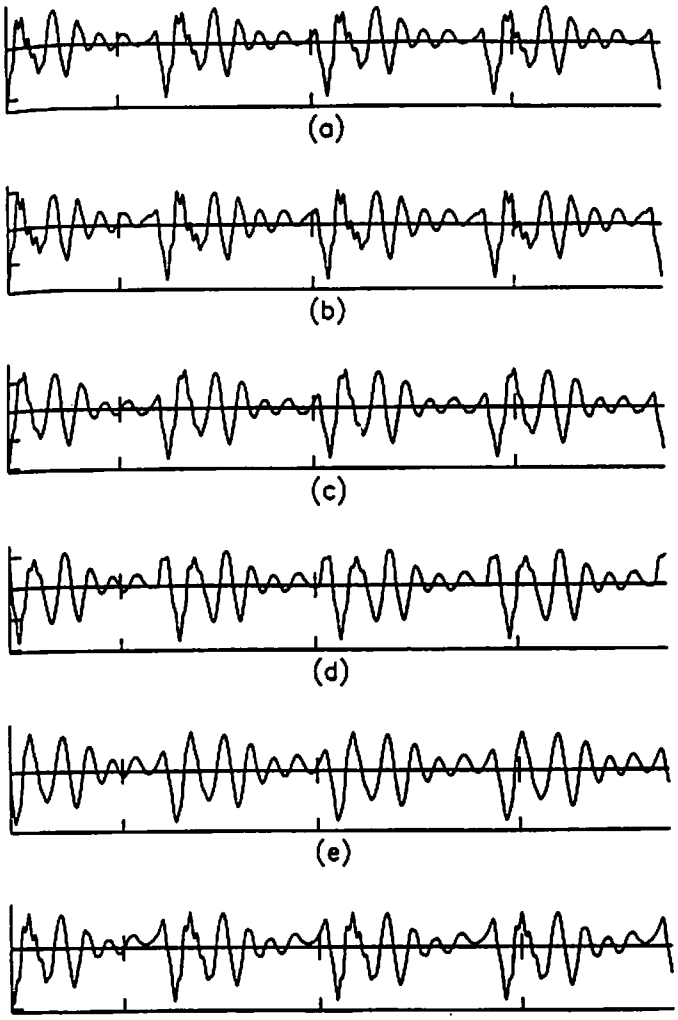


FIGURE 5. Recorded signals from microphones for stimulus of male subject M1 at 0° angle and 15 cm distance. Parts (a)–(f) correspond to microphones 1–6.

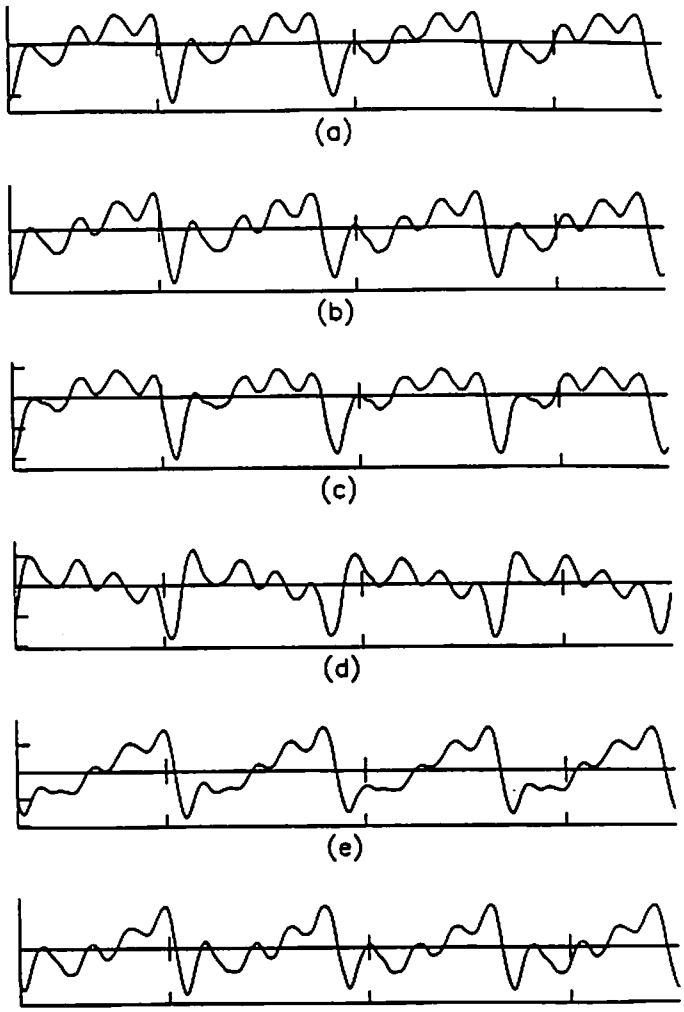


FIGURE 6. Recorded signals from microphones for stimulus of female subject F1 at 0° angle and 15-cm distance. Parts (a)–(f) correspond to microphones 1–6.

note that the condenser microphones (parts a, b, and f) show greater third-formant ripple on the positive peak following the largest negative peak (after vocal tract excitation). This suggests better mid-frequency response (around 2500 Hz) for the condenser microphones. For the female subject F1 phonating at 246 Hz (Figure 6), major waveform distortions are seen in parts (d) and (e), again for the dynamic microphones. Microphone 4 (part d) exhibits the only waveshape for which first-formant ripple drifts downward (instead of upward) after the negative peak. Microphone 5 (part e) attenuates even the first-formant ripple at this higher F_0.

Loss or distortion of formant ripple may affect the extraction of perturbation measures, especially if peak detection or waveform matching is used. This is because the waveform distortions may be different from cycle to cycle in the presence of vocal jitter and shimmer. For the waveform technique used in the present study, loss or distortion of formant ripple may be part of the reason why the female subject showed greater variability in the perturbation measures than the male subject. Careful examination of Figure 6 waveforms shows greater cycle-to-cycle variability than in Figure 5, which can best be seen by observing the events near zero crossings.

A quantitative comparison between the condenser microphones and the dynamic microphones for human phonations is shown in Figure 7. Recall that placement was at 15 cm, 0° angle. Scattergrams are shown for all amplitude perturbation measures against the means of the three microphones in each class. Part (a) is for the three condenser microphones and part (b) is for the three dynamic microphones. It appears that the scatter is about the same for the dynamic microphones as for the condenser microphones.

Figure 8 shows the same comparison for synthesized perturbations and modulations. The distance here was 4 cm and 0° angle. In groups of three, the condenser microphones perform somewhat better than the dynamic types for low perturbations (below 0.5%). Note how close microphone 2 is to the diagonal (asterisks in part a of the figure). This is by far the best microphone under the conditions we tested. Microphone 1, the omnidirectional professional-grade condenser microphone, deviates more from the diagonal, especially at low perturbations. Microphone 5, the consumer-grade dynamic microphone, shows the worst performance at low perturbations. If the consumer-grade microphones (5 and 6) are removed from the data set, condenser types perform

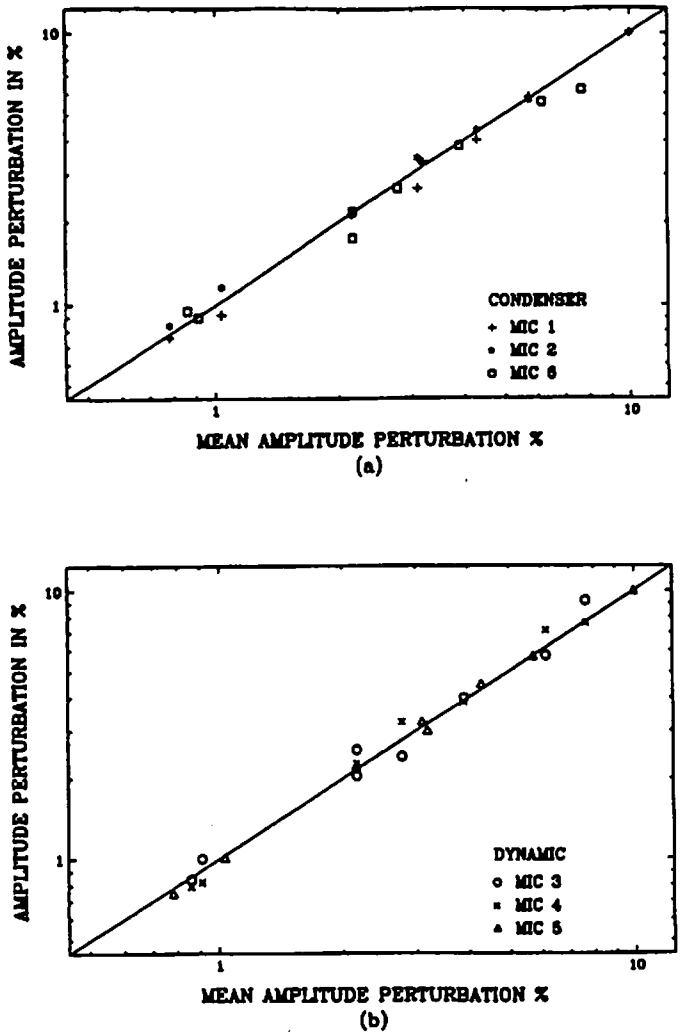


FIGURE 7. Scatter plot for amplitude perturbations (CV and P1) extracted from 4 subjects (2 male and 2 female) at 15 cm and 0° angle (a) condenser microphones, and (b) dynamic microphones.

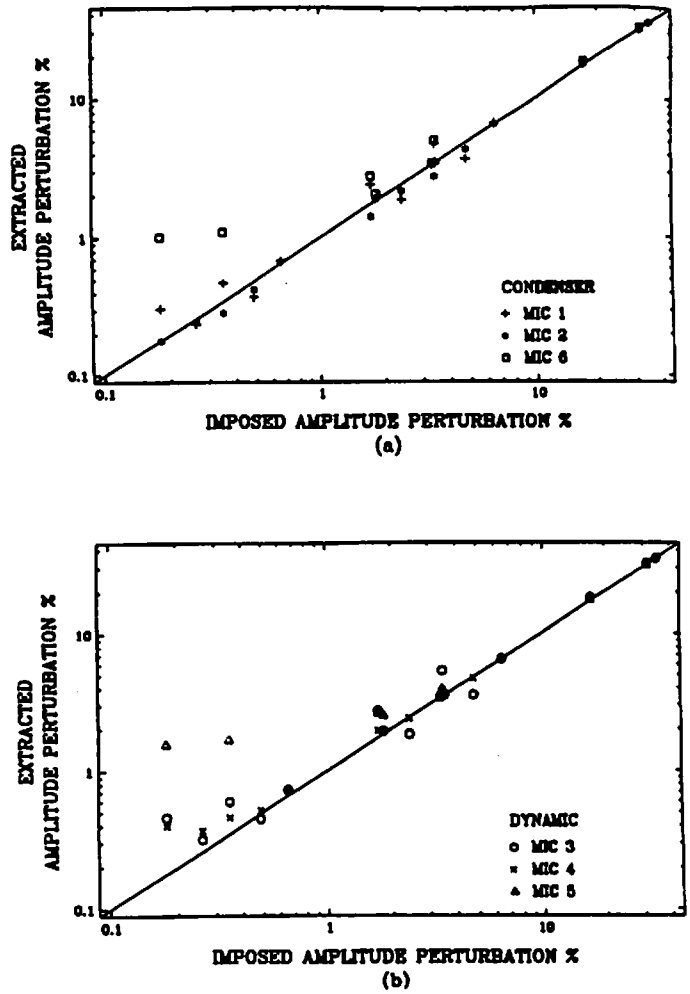


FIGURE 8. Scatter plot for amplitude perturbations (CV and P1) extracted from synthesized signals presented over a loudspeaker at 4 cm and 0° angle (a) condenser microphones, and (b) dynamic microphones.

collectively better than dynamic types. Note the deviations from the diagonal for all dynamic types when perturbations are under 0.5%. No individual dynamic microphone outperformed the best condenser microphone (number 2).

Microphone Distance

Table 6 presents an ensemble average over four professional microphones, two fundamental frequencies, and two measures (CV and P1) of perturbation to determine the baseline levels at different microphone positions. Note the general increase with distance, basically an order of magnitude from 4 cm to 1 m.

Scatter plots were made for both amplitude and frequency measures. These scatter plots relate the extracted perturbations to the imposed perturbations (direct measure). To conserve space, mainly the amplitude plots are shown because amplitude perturbation is most affected by distance and angle. Frequency plots were very similar in shape, however.

Figure 9 is a scatter plot of the amplitude perturbation measures (CV and P1) for the four professional-grade microphones at 4 cm. Fundamental frequencies of 100 Hz and 300 Hz are included and the angle is 0°. The plot is basically a combination of Figures 8(a) and (b), with consumer-grade microphones excluded. Note that microphone 2 (the cardioid

TABLE 6. Perturbation measures (ensemble averages over four professional microphones, two fundamental frequencies, and two measures CV and P1) for no imposed modulation with different microphone distances and angles to the source.

	Angle		
Distance	0	45	90
	Amplitude measures		
4 cm	0.24	0.31	0.30
30 cm	1.33	1.75	2.09
1 m	3.96	5.03	6.72
	Frequency measures		
4 cm	0.03	0.04	0.04
30 cm	0.16	0.23	0.31
1 m	0.49	0.78	1.28

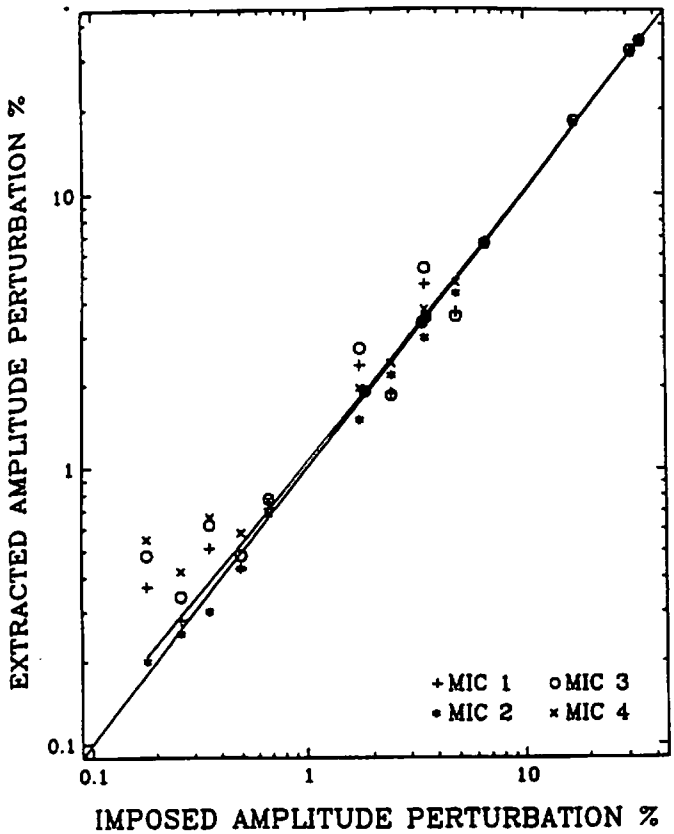


FIGURE 9. Scatter plot for amplitude perturbations from professional-grade microphones, all modulations, F_0 of 100 Hz and 300 Hz, 0° angle, and 4-cm distance.

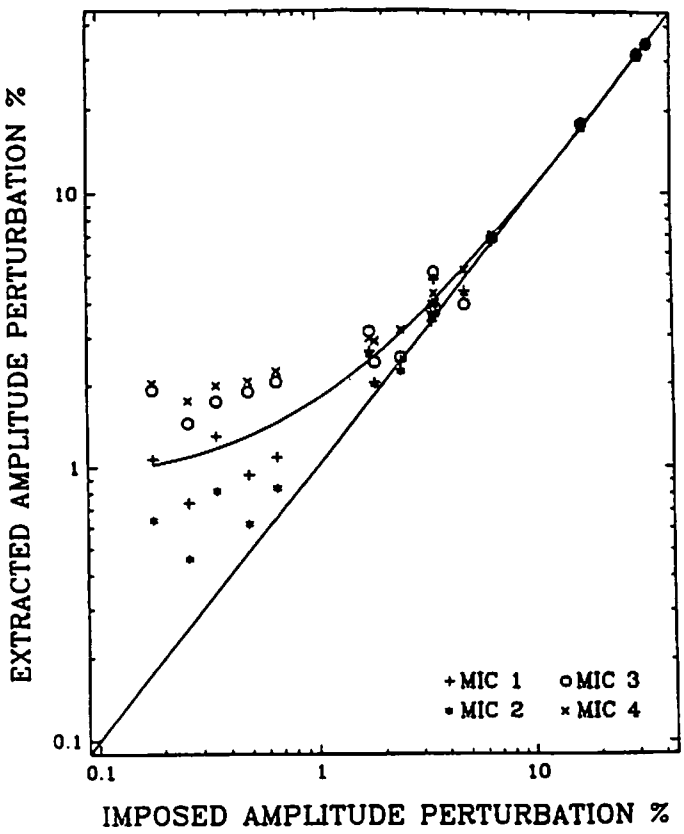


FIGURE 10. Scatter plot for amplitude perturbations from professional-grade microphones, all modulations, F_0 of 100 Hz and 300 Hz, 0° angle, and 30-cm distance.

condenser) shows the best results and microphone 3 (the omnidirectional dynamic) shows the worst results. For perturbations between 1% and 10%, the errors in extraction range between 1% and 10% at this distance. This suggests that the electroacoustic link contributes to extraction errors between 1 part in 10^4 (80 dB) and 1 part in 10^2 (40 dB), depending on the manner in which the signal is distorted. A slight upward trend from the diagonal at low perturbations (near 0.2%) is observed by fitting a second order polynomial to all the data.

Figure 10 is a similar plot at a distance of 30 cm, and Figure 11 at a distance of 1 m. In these scatter plots, the second-order fit shows a major trend. There are progressively larger deviations from the diagonal for greater distances. For a commonly used 30-cm distance, a 1% amplitude perturbation is inflated by a factor of 2, and a 0.1% perturbation by a factor of 10 or more. At 1-m distance, only amplitude perturbations of 10% or more have reasonable precision.

To bracket the frequency perturbation results and to show their similarity to amplitude perturbations, Figures 12 and 13 show scatter plots for frequency measures at 4 cm and 1 m, respectively. These should be compared with the amplitude measures in Figures 9 and 11.

The microphone rankings were retained at all of the distances and with all of the measures. Best results were generally obtained with microphone 2, second best with microphone 1, and worst with microphones 3 or 4. This is consistent with previous discussions on sensitivity and phase distortion of these microphones. Phase of the recorded signals was affected by distance, as observed informally by comparing waveforms at various distances and angles. This is a topic for another discussion, however, because it involves details of sound field patterns and room acoustics.

Microphone Angle

Angular orientation of the microphone had an effect on the measures primarily at large distances (Table 6). At a distance of 4 cm, the measures were inflated by 20%–30% for angles of 45° and 90°. At a distance of 30 cm, the measures were about 50% larger at 45° and approached 100% inflation at 90°. At 1-m distance, the inflation with angle was slightly greater than at 30 cm. A typical scatter plot for amplitude measures is shown in Figure 14 for an angle of 45° and a distance of 30 cm. This plot should be compared with Figure 10 to get the graphic effect of angle.

It is interesting to ask whether microphone preference due to directionality (cardioid versus omnidirectional) changes with angular orientation. Our data showed that it does not. In every case tested [3 distances × 3 angles], the cardioid condenser microphone performed best. The radiation pattern from the loudspeaker and the reception pattern from the cardioid microphone are apparently in the 0–90° range. In other words, the cardioid microphone gains sensitivity for

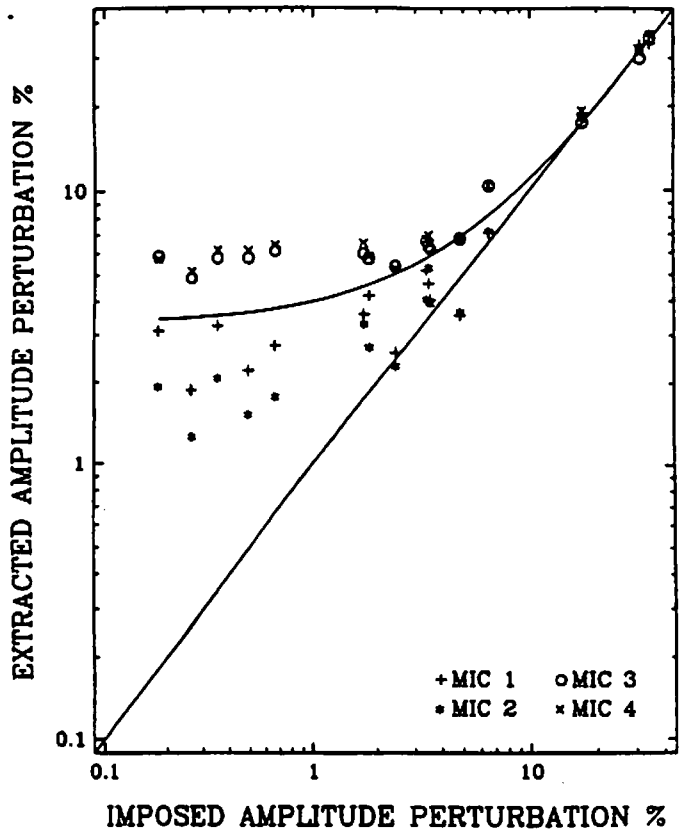


FIGURE 11. Scatter plot for amplitude perturbations from professional-grade microphones, all modulations, F_0 of 100 Hz and 300 Hz, 0° angle, and 1-m distance.

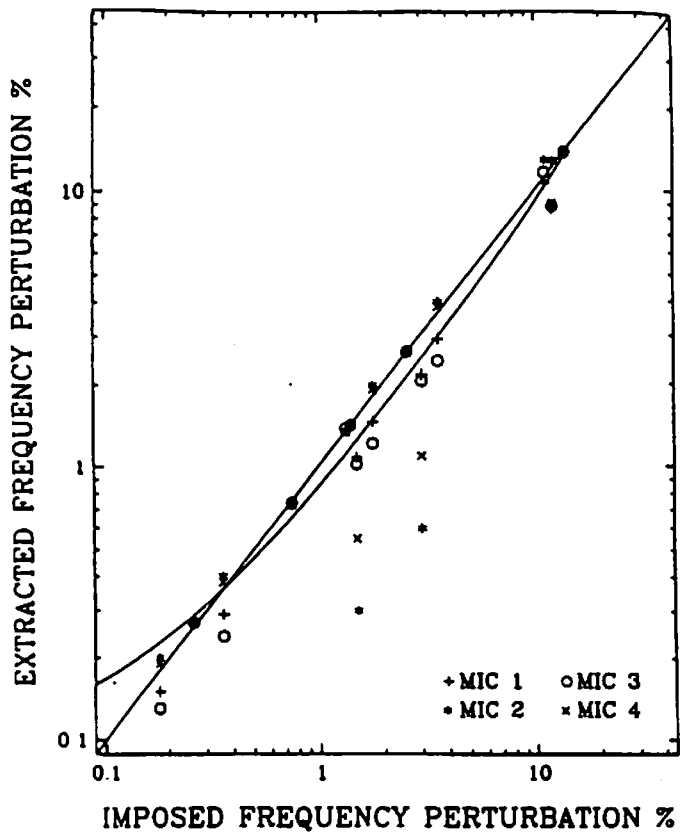


FIGURE 12. Scatter plot for frequency perturbations from all professional grade microphones, all modulations, F_0 of 100 Hz and 300 Hz, 0° angle, and 4-cm distance.

near 0° placement, but does not lose much sensitivity anywhere in the first quadrant.

Conclusions

Microphone sensitivity has the greatest effect on the extraction of perturbation measures from recorded voice signals. When sensitivity is low, the signal can be preamplified, but noise is amplified along with the signals. In this study, sensitivity was the largest variable between the professional-grade microphones, with the condenser type being approximately 20 dB more sensitive than the dynamic type.

Baseline frequency perturbations with professional grade microphones and digital recording equipment were about 0.05%, and baseline amplitude perturbations were about 0.3% at a distance of 4 cm. These are the values of jitter and shimmer when there are no imposed modulations and when the entire electroacoustic link affects the signal (digital-to-analog converter, power amplifier, loudspeaker, microphone, preamplifier filter, DAT recorder, and analog-to-digital converter). Some consumer-grade microphones used in conjunction with the same equipment and analysis programs inflated the frequency perturbation to a range of 0.1–0.2% and amplitude perturbation to a range of 1.0–2.0%.

When the microphone distance was changed from 4 cm to 1 m, perturbation measures increased by an order of magnitude. This increase can be explained by a loss in signal-to-noise ratio. In this experiment, a large increase in voltage gain (microphone preamplifier) was necessary to overcome the inverse square loss of acoustic power by distance. This makes it virtually impossible, in a typical sound isolation booth, to resolve small jitter and shimmer values (those in the 0.1 to 0.5% range) at distances greater than a few cm from the mouth.

Another variable between microphone type was phase distortion. This may have significant effects on the extraction algorithms for perturbation, especially those that depend on specific events on the waveform (peaks, zero crossings, etc.), posing a problem for certain types of analyses. The interactions among distance, angle, and harmonic structure of the signal complicate the transduction process, which in turn introduces variability into the perturbation analysis. Once phase distortion has altered the microphone signal, from whatever source, it may not be possible to obtain meaningful voice perturbation values. An exception might be the case where the distortion is linear and known, in which case a corrective inverse distortion can be applied by a pre-processor.

In this study there appeared to be some evidence that the cardioid pattern is preferable to the omnidirectional pattern, probably because of its greater on-axis gain. But this is based on a very limited sampling of microphones, making generalizations impossible at this stage. At great distances and wide angles, there may be more of a pattern effect that would ultimately favor the omnidirectional microphone, but this would need further study.

Under some conditions (e.g., clinical applications where the patient is immobile), microphone placement may be

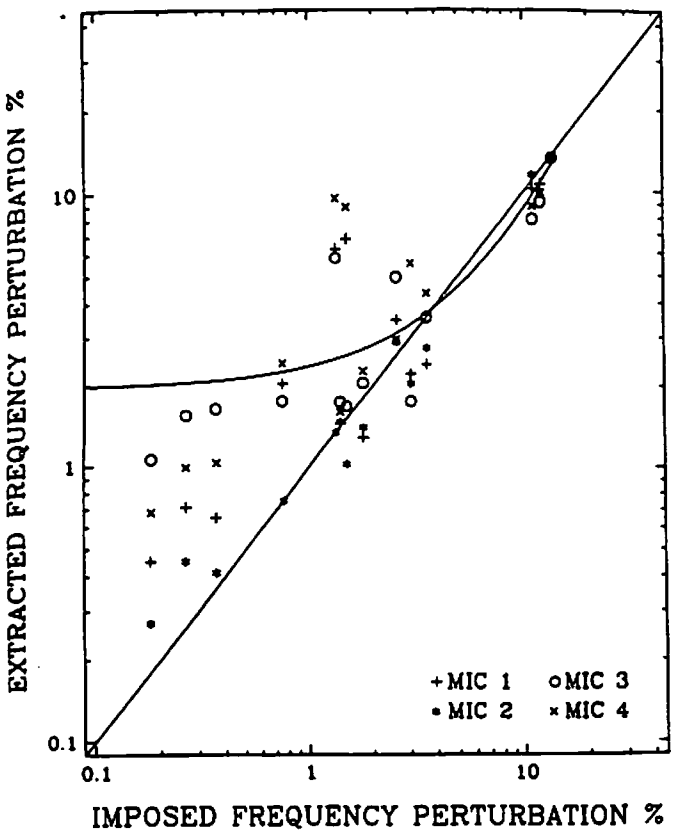


FIGURE 13. Scatter plot for frequency perturbations from all professional-grade microphones, all modulations, F_0 of 100 Hz and 300 Hz, 0° angle, and 1-m distance.

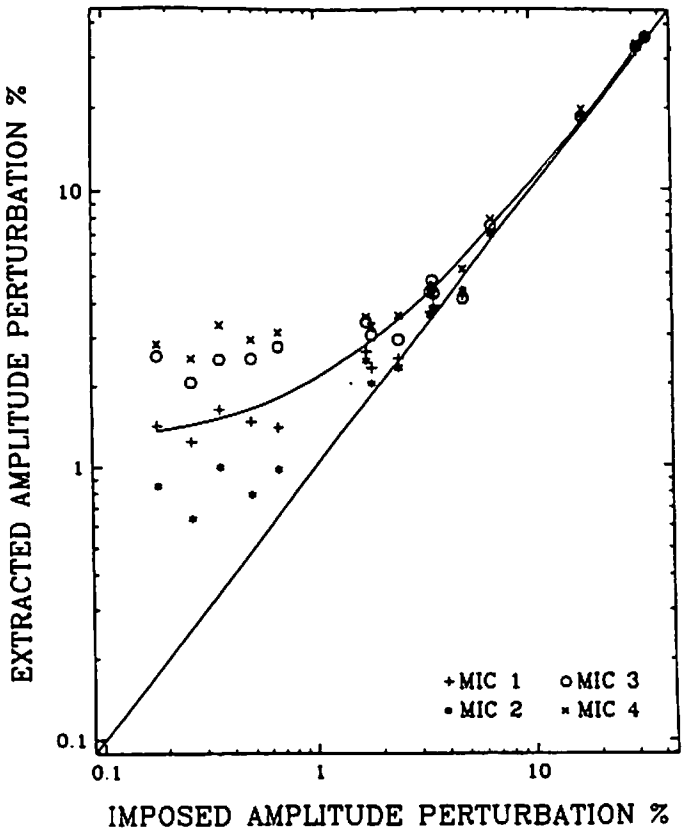


FIGURE 14. Scatter plot for amplitude perturbations from professional-grade microphones, all modulations, F_0 of 100 Hz and 300 Hz, 45° angle, and 30-cm distance.

restricted to distances greater than a few cm, or a relatively noisy environment. A future study is needed to determine the effects of room reverberation and ambient noise. Judging by the current results, however, severe limitations are expected.

It is also important to note that, in an attempt to simplify this study, only sustained phonation was used for the stimulus. For running speech there are considerable aerodynamic artifacts (plosives, DC drift, etc.) that will affect the acoustic signal at distances near the mouth. Therefore, a compromise in microphone placement is needed, depending upon the type of vocal tasks that are used.

If a recommendation had to be given from this initial study, such a compromise would be to place a professional-grade cardioid or omnidirectional condenser microphone a few centimeters from the mouth, at 45° to 90°. A miniaturized head-mounted condenser microphone is presently under consideration. Microphone sensitivity should be no less than −60 dB, and a high-pass preconditioning filter may be necessary to reduce aerodynamic artifact and low-frequency acoustic noise for near-mouth placement (in addition to the off-axis orientation). The preconditioning filter should be linear phase. Digital finite impulse response (FIR) filters can be used if free-standing filters are not available. Low-pass anti-aliasing filters, typically set above 5 kHz, need not be linear phase because they do little to distort the waveshape. The recommendations apply only to perturbation analysis on sustained vowels at this stage.

Acknowledgment

This study was supported by a grant from the National Institutes of Health, Grant No. R01 DC00387-04. The authors wish to thank Martin Rothenberg and Ron Scherer for technical assistance, Chwen-Geng Guo, Larry Brown, and Mitch Wolfe for assistance in computer analysis, and Pamela Rios and Julie Lemke for assistance in manuscript preparation.

References

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

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

Horii, Y., & Hata, K. (1988). A note on phase relationships between frequency and amplitude modulation in vocal vibrato. *Folia Phoniatrica, 40,* 303–311.

House, A. (1960). A note on optimal vocal frequency. *Journal of Speech and Hearing Research, 2*(1), 55–60.

Karnell, M. P. (1992). Laryngeal perturbation analysis: Minimum length of analysis window. *Journal of Speech and Hearing Research, 34,* 544–548.

Karnell, M. P., Scherer, R. S., & Fischer, L. B. (1991). Comparison of acoustic voice perturbation measures among three independent voice laboratories. *Journal of Speech and Hearing Research, 34,* 781–790.

Milenkovic, P. (1987). Least mean square measures of voice perturbation. *Journal of Speech and Hearing Research, 30,* 529–538.

Pinto, N., & Titze, I. (1990). Unification of perturbation measures in speech analysis. *Journal of the Acoustical Society of America, 87*(3), 1278–1289.

Scherer, R. C., Gould, W. J., Titze, I. R., Meyers, A. D., & Sataloff, R. T. (1988). Preliminary evaluation of selected acoustic and glottographic measures for clinical phonatory function analysis. *Journal of Voice, 3*, 230–244.

Schutte, H., & Seidner, W. (1983). Recommendation by the Union of European Phoniatricians (UEP): Standardizing voice area measurement/phonetography. *Folia Phoniatrica, 35*, 286–288.

Titze, I. R., & Liang, H. (1993). Comparison of F_0 extraction methods for high-precision voice perturbation measurements. *Journal of Speech and Hearing Research, 36*, 1120–1133.

Winholtz, W. S., & Ramig, L. O. (1992). Vocal tremor analysis with the vocal demodulator. *Journal of Speech and Hearing Research, 35*, 562–573.

Received December 28, 1992
Accepted July 2, 1993

Contact author: Ingo R. Titze, National Center for Voice and Speech, Department of Speech Pathology and Audiology, The University of Iowa, 330 WJSHC, Iowa City, IA 52242-1012.

Statement of Ownership, Management and Circulation
(Required by 39 U.S.C. 3685)

1A. Title of Publication	1B. PUBLICATION NO.	2. Date of Filing
Journal of Speech and Hearing Research (JSHR)	0 2 2 - 4 6 8 5	Oct. 1, 1993
3. Frequency of Issue: Bimonthly	3A. No. of Issues Published Annually: Six issues	3B. Annual Subscription Price: $114 (US) $126 (Foreign)

4. Complete Mailing Address of Known Office of Publication (Street, City, County, State and ZIP+4 Code) (Not printers)
American Speech-Language-Hearing Association, 10801 Rockville Pike, MD 20852-3279

5. Complete Mailing Address of the Headquarters or General Business Offices of the Publisher (Not printer)
Same as #4 above.

6. Full Names and Complete Mailing Address of Publisher, Editor, and Managing Editor (This item MUST NOT be blank)
Publisher (Name and Complete Mailing Address)
American Speech-Language-Hearing Association, 10801 Rockville Pike, Rockville, MD 20852-3279
Editor (Name and Complete Mailing Address)
John H. Saxman, Ph.D.
Box 146, Teachers College, Columbia University, New York, NY 10027
Managing Editor (Name and Complete Mailing Address)

7. Owner (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given. If the publication is published by a nonprofit organization, its name and address must be stated.) (Item must be completed.)

Full Name	Complete Mailing Address
American Speech-Language-Hearing Association (a corporation organized not for profit)	10801 Rockville Pike, Rockville, MD 20852-3279

8. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages or Other Securities (If there are none, so state)

Full Name	Complete Mailing Address
Crestar Bank N.A.	1445 New York Ave., NW, Washington, DC 20005
American National Insurance Co. Mortgage and Real Estate Investments	1 Moody Plaza, Galveston, TX 77550

9. For Completion by Nonprofit Organizations Authorized To Mail at Special Rates (DMM Section 424.12 only)
The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes (Check one)
(1) [X] Has Not Changed During Preceding 12 Months (2) [ ] Has Changed During Preceding 12 Months (If changed, publisher must submit explanation of change with this statement.)

10. Extent and Nature of Circulation (See instructions on reverse side)	Average No. Copies Each Issue During Preceding 12 Months	Actual No. Copies of Single Issue Published Nearest to Filing Date
A. Total No. Copies (Net Press Run)	41,250	40,700
B. Paid and/or Requested Circulation 1. Sales through dealers and carriers, street vendors and counter sales	---	---
2. Mail Subscription (Paid and/or requested)	37,974	37,595
C. Total Paid and/or Requested Circulation (Sum of 10B1 and 10B2)	37,974	37,595
D. Free Distribution by Mail, Carrier or Other Means Samples, Complimentary, and Other Free Copies	26	34
E. Total Distribution (Sum of C and D)	38,000	37,629
F. Copies Not Distributed 1. Office use, left over, unaccounted, spoiled after printing	2,250	3,071
2. Return from News Agents	---	---
G. TOTAL (Sum of E, F1 and 2—should equal net press run shown in A)	41,250	40,700

11. I certify that the statements made by me above are correct and complete
Signature and Title of Editor, Publisher, Business Manager, or Owner
Frederick T. Spahr, Ph.D., Executive Director

PS Form 3526, January 1991 (See instructions on reverse)

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.

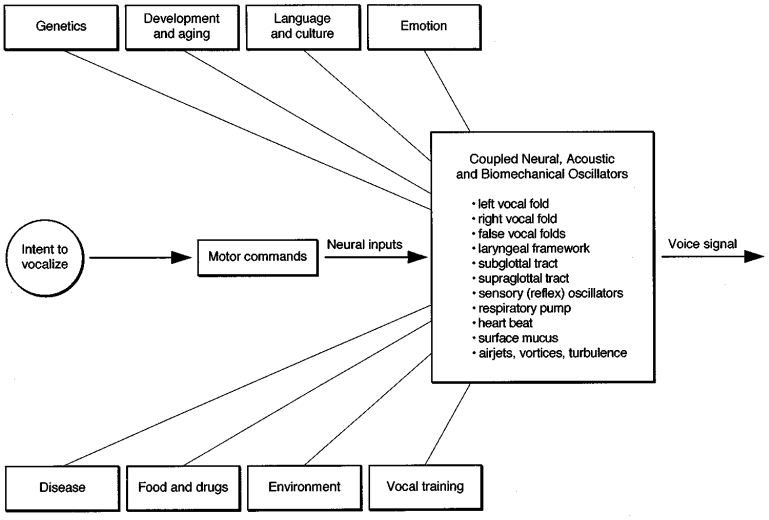


Figure 1.
A list of biological oscillators involved in voice production and factors that may influence them.

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

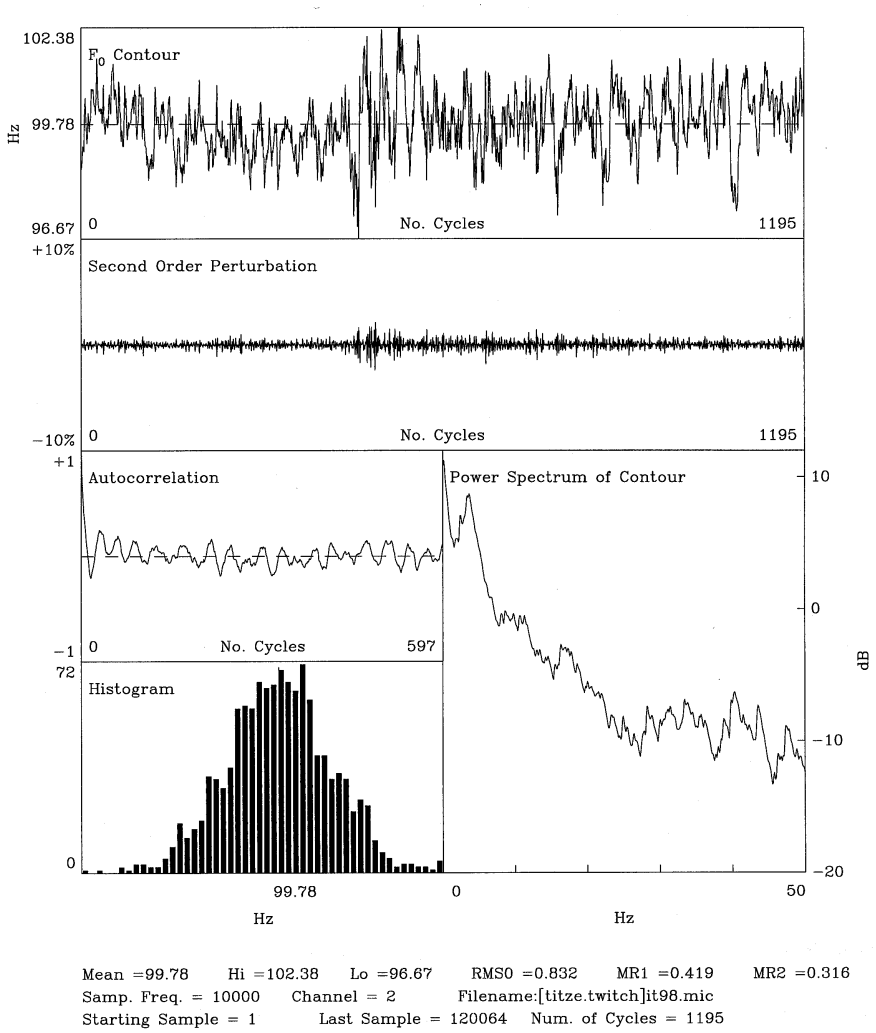


Figure 2. A fundamental frequency (F_o) profile used for perturbation analysis. The subject was a normal adult male phonating a steady [b] vowel at approximately 100 Hz for about 12 seconds.

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*.

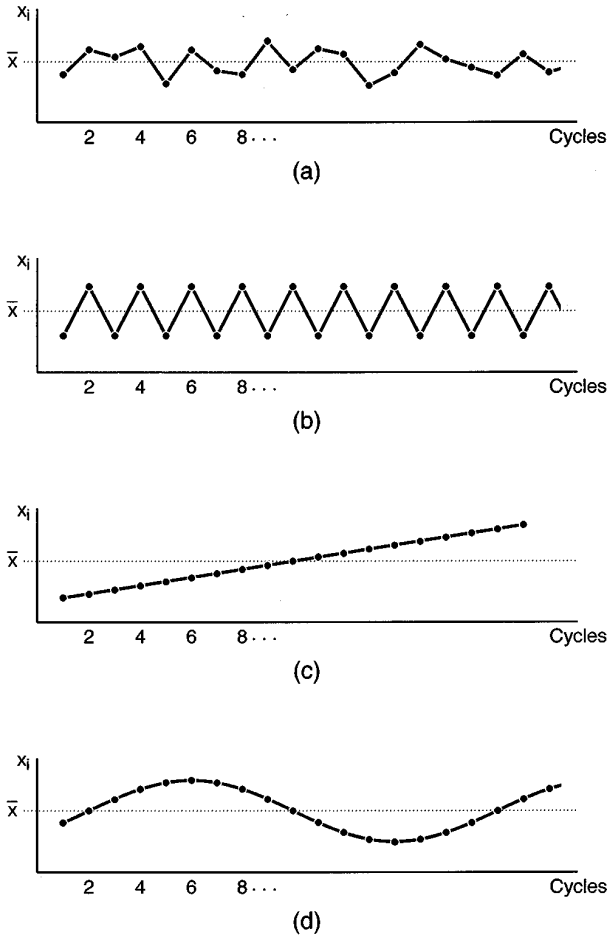


Figure 3. Modulations of a cyclic parameter x_i around the mean value (a) random, (b) alternating, (c) linear trend, and (d) sinusoidal.

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).

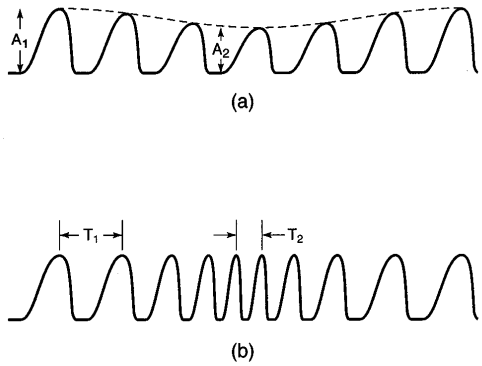


Figure 4. (a) Amplitude modulation (AM), (b) frequency modulation (FM) of a series of glottal pulses.

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*).

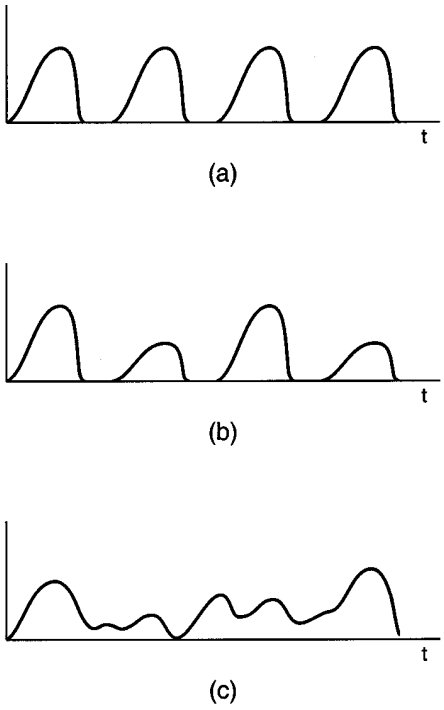


Figure 5. A series of glottal pulses showing evidence of bifurcation. (a) periodic vibration, (b) period doubling, (c) chaotic vibration.

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.

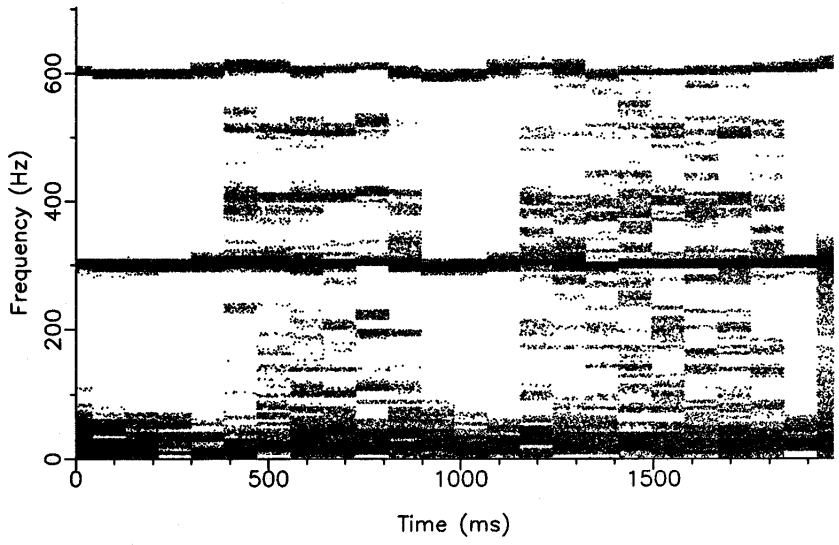


Figure 6. Narrow-band computer spectrogram for a patient with hyperfunctional childhood dysphonia. Abrupt transitions to different phonatory regimes are visible, indicating bifurcations in vocal fold vibration.

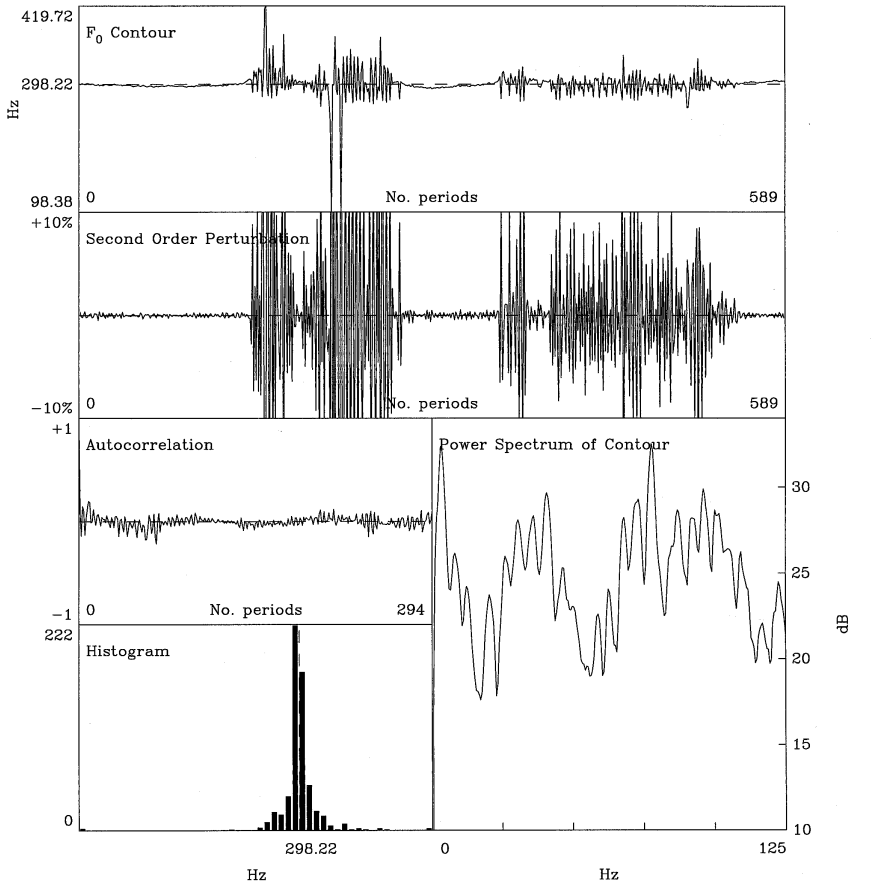


Figure 7. Fundamental frequency (F_o) profile for the patient with hyperfunctional childhood dysphonia.

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.

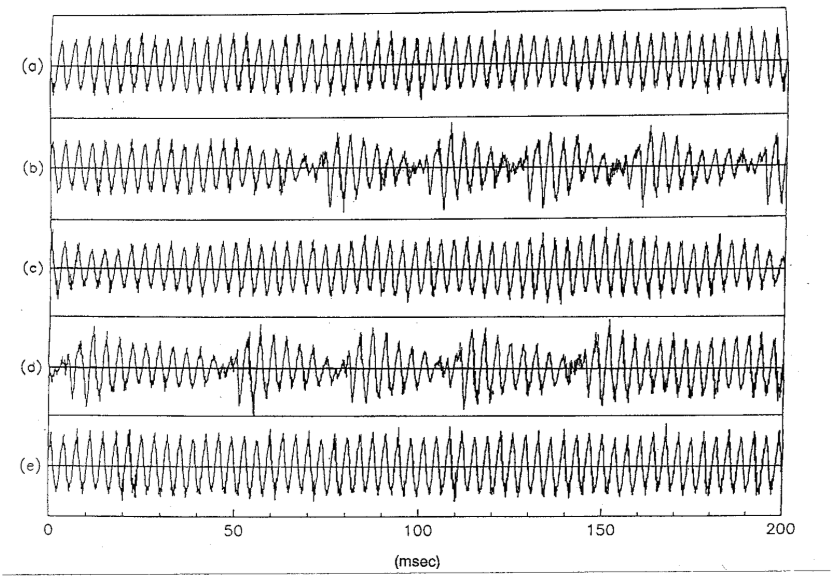


Figure 8. Microphone signal of a patient with unilateral lyarngeal nerve paralysis. Parts (a) to (e) should be viewed serially, 200 ms per segment, for a total of 1s (After Herzel et al, 1994).

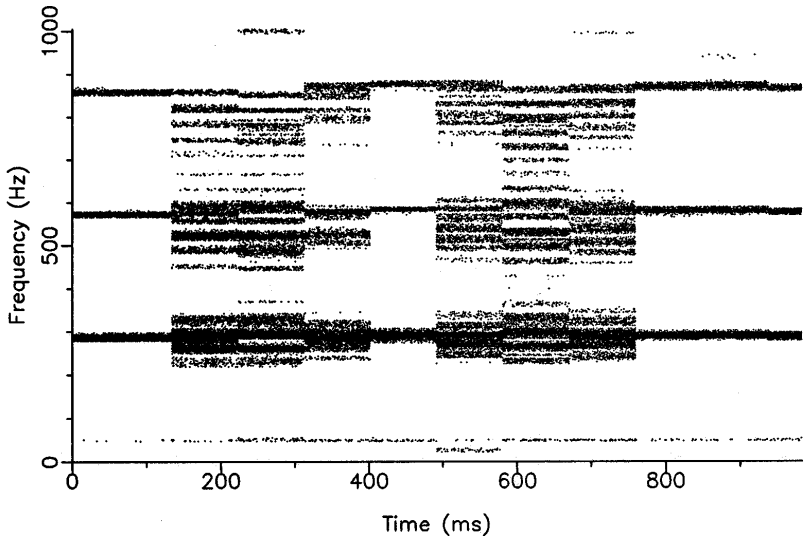


Figure 9. Narrow-band spectrogram of a patient with unilateral laryngeal nerve paralysis, corresponding to the waveform in Figure 8.

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.

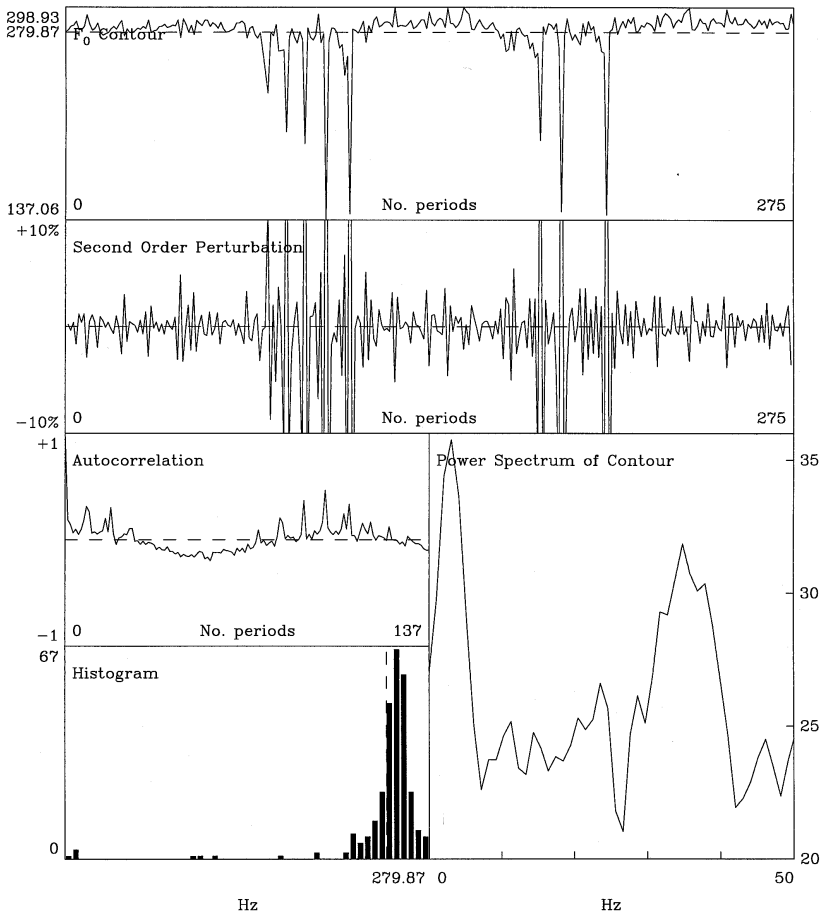
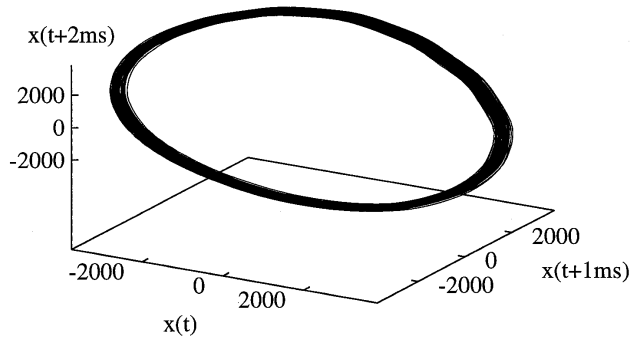
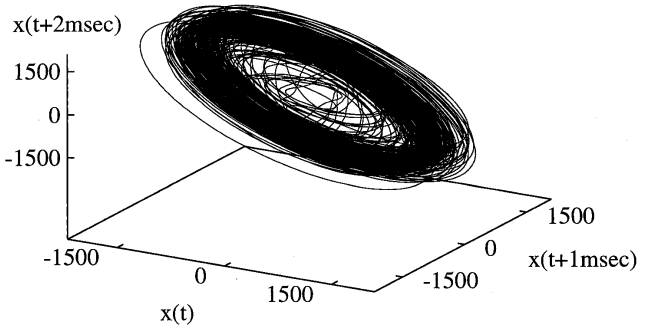


Figure 10. Fundamental frequency (F_o) profile for the patient with unliateral laryngeal nerve paralysis, corresponding to Figures 8 and 9.

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.