A Nonlinear Operator-Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment

John H. L. Hansen,* *Senior Member, IEEE*, Liliana Gavidia-Ceballos, and James F. Kaiser, *Fellow, IEEE*

***Abstract*— Traditional speech processing methods for laryngeal pathology assessment assume linear speech production with measures derived from an estimated glottal flow waveform. They normally require the speaker to achieve complete glottal closure, which for many vocal fold pathologies cannot be accomplished. To address this issue, a nonlinear signal processing approach is proposed which does not require direct glottal flow waveform estimation. This technique is motivated by earlier studies of airflow characterization for human speech production. The proposed nonlinear approach employs a differential Teager energy operator and the energy separation algorithm to obtain formant AM and FM modulations from filtered speech recordings. A new speech measure is proposed based on parameterization of the autocorrelation envelope of the AM response. This approach is shown to achieve impressive detection performance for a set of muscular tension dysphonias. Unlike flow characterization using numerical solutions of Navier–Stokes equations, this method is extremely computationally attractive, requiring only a small time window of speech samples. The new noninvasive method shows that a fast, effective digital speech processing technique can be developed for vocal fold pathology assessment without the need for direct glottal flow estimation or complete glottal closure by the speaker. The proposed method also confirms that alternative nonlinear methods can begin to address the limitations of previous linear approaches for speech pathology assessment.**

***Index Terms*— Nonlinear acoustics, speech analysis, speech pathology, vocal system.**

I. Introduction

The presence of vocal fold pathology can cause significant changes in the normal vibratory patterns of the vocal folds, which in turn impacts the resulting quality of speech production. Early signs of vocal fold malfunction are normally associated with breathiness or hoarseness in the resulting speech signal. In recent years, researchers in laryngology and speech pathology have become increasingly interested in acoustic analysis of normal and pathological voices [10], [17], [18], [30]. One reason for this trend is that acoustic methods have the potential to provide quantitative techniques for clinical assessment of laryngeal and vocal tract function. Though several methods currently exist for laryngeal and vocal tract research and diagnosis (e.g., laryngoscopy, glottography, electromyography, stroboscopy, and acoustic analysis) [39]; acoustic analysis appears to have an advantage over other methods because of its nonintrusive nature and its potential for providing quantitative data with reasonable analysis time.

Previous signal processing techniques have concentrated on acoustic signal analysis, where methods have attempted to identify the acoustic parameter changes due to pathology across time. Examples of these studies include Feijoo and Hernandez [13], Hirano *et al.* [19], Hoyt *et al.* [20], and Kasuya *et al.* [23]. A number of different acoustic measures have resulted from such studies including tone period perturbation (jitter), amplitude perturbation (shimmer), vocal noise, and measures based on spectral or formant analysis. Modest performance in terms of error percentages obtained for healthy versus pathology classification using acoustical analysis, limits its use in a clinical environment (i.e., error rates between 10–20%).

Other methods have focused on estimating the glottal flow waveform based on inverse filtering [16], which attempt to remove vocal tract and lip radiation characteristics, resulting in an estimated glottal flow waveform [1], [41], [42]. It has been suggested that the glottal flow waveform represents the overall sound pressure wave emanating from the glottis. Previous approaches to characterizing vocal fold pathology suggest that it is first necessary to estimate the glottal flow waveform, and that subsequent analysis on this waveform should reveal changes in speech quality, or the presence of an irregular mucosal wave due to pathology.

In science, linear approximations are often used to model nonlinear phenomena. As an example, voltage–current characteristics of transistors have an operating range which is quasilinear. If the transistor is kept in this linear range, the linear model performs well. A linear model is often also easier to estimate and adapt for time-varying systems. The popular model for characterizing speech production is based on linear acoustic theory (for a review, see [12], [14], [31], [34]). In this linear source/filter model, linear acoustic tube theory is used to derive a digital filter representation for speech production assuming planar sound propagation [see Fig. 1(a)].

Manuscript received November 10, 1995; revised May 30, 1997. This work was supported by a grant from The Whitaker Foundation. *Asterisk indicates corresponding author.*

*J. H. L. Hansen is with the Departments of Electrical and Biomedical Engineering, Duke University, P.O. Box 90281, Durham, NC 27708-0281 USA. He is also with the Robust Speech Processing Laboratory, Department of Electrical and Computer Engineering, Duke University, P.O. Box 90291, Durham, NC 27708-0291 USA (e-mail: jhlh@ee.duke.edu; htp://www.ee.duke.edu/research/speech).

L. Gavidia-Ceballos was with the Department of Biomedical Engineering, Duke University, Durham, NC 228708-0281 USA. She is now with the Universidad Simon Bolivar, Caracas, Venezuela.

J. F. Kaiser is with the Robust Speech Processing Laboratory, Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708-0291.

Publisher Item Identifier S 0018-9294(98)01606-1.

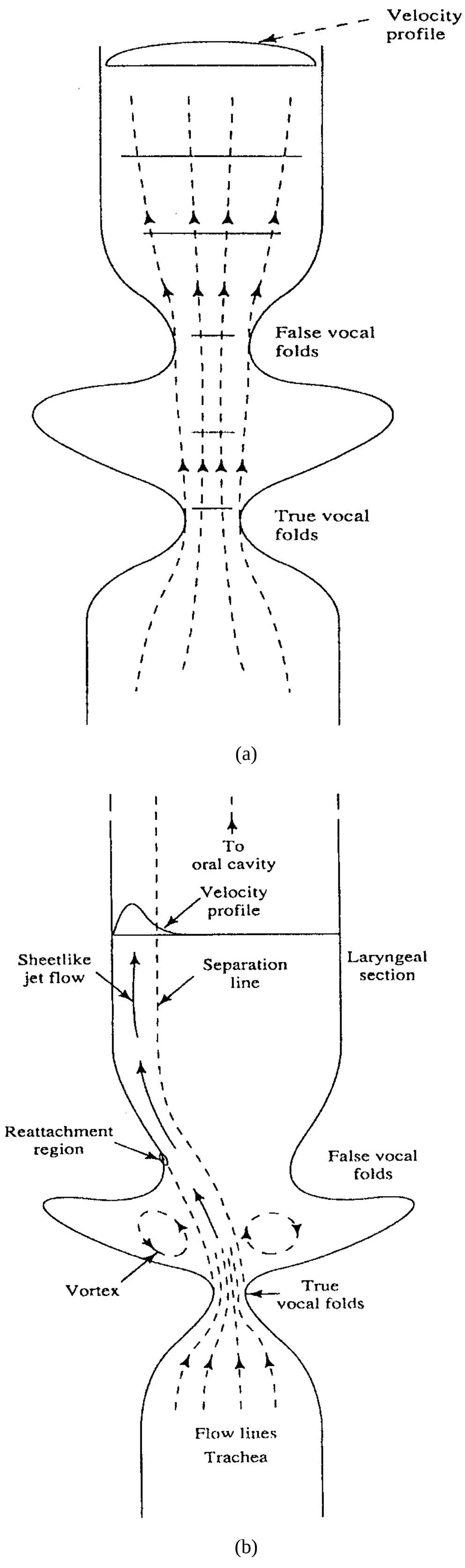

Fig. 1. The (a) classical interpretation of airflow propagation along the vocal system and (b) a nonlinear fluid dynamics interpretation of airflow propagation along the vocal tract for speech production [12], [20b].

While this model has served the speech community well in the areas of coding and recognition, there is evidence that speech production incorporates nonlinear traits [36]–[38]. In fact, the physics of the larynx suggests that idealized planar propagation as illustrated in Fig. 1(a) cannot occur. Sound pressure and volume velocity measurements within cast models of the human speech system suggest that the nonlinear fluid dynamics as illustrated in Fig. 1(b) to be a more realistic means of characterizing sound pressure along the vocal tract. Studies by Teager [36], [37] suggest that the vorticies located in the vicinity of the false vocal folds provide the necessary source of excitation during the closed phase of the vocal folds. It is also noted that the propagating sound waves follow lines of airflow, and that high-speed flows tend to adhere to the vocal-tract walls. A complete solution for such a nonlinear fluid dynamic model requires the solution of the Navier–Stokes equations and has been achieved in a very limited sense for a stationary phoneme [38]; however, at present it is not computationally feasible to obtain a Navier–Stokes solution for a time-varying speech model, the main problem being the difficulty in obtaining reliable boundary information.

Although modeling actual speech production using fluid flow characteristics would be useful, the ability to characterize fluid flow properties in actual vocal fold pathology patients and maintaining a noninvasive procedure would be difficult. Instead, it is suggested that features derived using a nonlinear speech framework could reveal the potential of alternative speech models to the traditional linear approach. The objective, therefore, is to formulate a nonlinear processing technique that can extract further information from the speech signal, including nonlinear excitation sources which are hypothesized to play a crucial role in both healthy and nonhealthy speech production.

A previous application of nonlinear energy principles for speech pathology was considered by Cairns *et al*. [4]. In that study, an algorithm was formulated using Teager energy profiles from bandpass and lowpass filtered speech under normal and hypernasal pathology. The method capitalized on the hypothesis that hypernasal speech contains antiformants and nasal formants, which are absent from normal speech. A distinct change in the Teager energy profiles was observed between neutral and hypernasal speech, and used to formulate a successful hypernasality detection algorithm. The focus of this previous study was pathology assessment in the vocal tract. No studies have yet considered such a nonlinear approach for vocal fold pathology assessment.

II. Background: Laryngeal Speech Pathology

Perceptive voice quality varies depending on the level and extent of glottal opening/closing (phonation types). Certain laryngeal pathologies impede the vocal folds from fully closing during glottal vibration. The main factors which govern vocal fold vibration are [6]: 1) vocal fold position, or the extent to which the vocal folds are adducted/abducted; 2) vocal fold myoelasticity, or the degree of elasticity of the vocal folds, (determined by the position and degree of tension as determined by muscle contraction); 3) the extent of the air pressure drop across the folds. For normal voice production, the vocal folds are relatively relaxed with a periodic opening and closing movement (i.e., roughly 50% duty-cycle in open/close phase). Several nonneutral phonation types exist and can indicate the prospect of more severe problems in speech production. These include whisper phonation, breathy/murmur phonation, creaky voice, vocal fry, and glottal blow. Incomplete glottal closure can arise from 1) paralysis or injury to one or both folds, 2) asymmetry between the two folds resulting from a laryngeal growth, or 3) swelling of the vocal folds. Incomplete glottal closure may also result from considerably different muscle tension or length of the vocal folds [25]. Asymmetry may further cause diplophonia (two tones), because the two folds do not vibrate in unison, resulting in a vocal tract which is excited at two different fundamental frequencies. Finally, vocal fold nodules, polyps, papillomas, contact ulcers, carcinoma, and recurrent laryngeal nerve paralysis constitute examples of laryngeal pathologies that involve disturbance of normal vocal fold opening and closing.

A. Functional Voice Disorders

A voice disorder is termed "functional" if it is primarily due to misuse or abuse of the anatomically and physiologically intact vocal system. These disorders are often difficult to detect with laryngoscopy since the larynx appears to be normal, though there are signs of laryngeal muscle tension and poor voice quality. The cause of functional disorders may be obvious or obscure, but prolonged abuse may result in the development of ulcers, polyps, nodules, or granulomas of the vocal cords. Koufman and Blalock [26], [27] classified functional disorders into five discrete groups: 1) hysterical dysphonia, 2) habituated hoarseness, 3) falsetto, 4) voice abuse, and 5) postoperative dysphonia. Each functional type possesses its own characteristics of onset and voice quality. For example, hysterical dysphonia is associated with a sudden loss of voice which is stable and "pitch-locked" (if voice is present). Habituated hoarseness is persistent hoarseness usually following an acute episode of laryngitis. Voice quality is breathy and/or raspy, and also shows signs of being pitch-locked. Falsetto is a developmental or sudden onset of abnormally high-pitched (locked) speech. Vocal abuse results from 1) overuse, 2) forcing pitch to be abnormally low, and 3) abnormal muscular tension in the larynx. Vocal abuse can also be associated with "secondary" vocal fold nodules, polyps, ulcers, and granulomas. Functional Types (1–3) are all stable dysfunctions which are not intermittent or fluctuant [26]. Previous clinical studies suggest that observable features of voice quality from analysis and characterization of a reliable glottal airflow waveform would be desirable for identification of these disorders. The most common functional voice disorder is the tension–fatigue syndrome. This typically occurs in nonprofessional voice users and is characterized by poor breath control, excessive muscle tension, dysphonia, and a marked restriction of the vocal range [32]. Types 4) and 5), which can be intermittent or fluctuant, are also important to consider since prolonged use can lead to more serious vocal system disorders.

Voice therapy techniques most commonly applied to patients with functional voice disorders are described in Prater [32]. Voice therapy requires identification of the primary cause of the disorder which may be due to vocal abuse and misuse. Treatment requires the patient to be made aware of situations where misuse occurs, so that this behavior can be modified or corrected. Voice therapy also includes techniques to achieve adequate breath support, as well as methods to soften a hard glottal attack.

B. Previous Laryngeal Assessment Methods

There are several routine procedures for laryngeal examination for clinical or research purposes. However, several limitations exist which include: 1) the larynx is located out of view, deep in the neck; 2) the interior of the larynx is dark, and must be adequately illuminated to allow for examination; 3) the movements of the vocal folds during phonation are too rapid to be captured by any conventional optical system. Despite these limitations, substantial efforts have resulted in a number of successful examination procedures [43]. Some of these include transnasal fiberoptic laryngoscopy (TFL), a procedure best suited for evaluation of vocal fold muscle tension during connected speech, since it interferes minimally with the supraglottic portion of the vocal tract. A more traditional method, such as mirror laryngoscopy enables the examiner to view the larynx without size or color distortion, which is often a problem with most optical-based instruments. Telescopic examination with stroboscopy is used primarily to evaluate the vocal folds for lesions, whereas TFL is used primarily to evaluate laryngeal biomechanics associated with speech production. Certain parameters extracted from electroglottographic (EGG) recordings, such as the closing rate and closed percentage, are particularly useful in assessing the degree of vocal fold tension, muscle stiffness or weakness, and degree of breathiness. Low closed percentage values indicate breathiness, whereas high values indicate pressed voice. However, the reliance on these acoustic measures remain imprecise for clinical differential diagnosis. A more detailed discussion on the different clinical procedures of laryngeal evaluation can be found in [3], [5]–[7], [11], [18], [24], and [25]. In the next section, we discuss the basic principles of the Teager Energy Operator (TEO) and its use in nonlinear speech production analysis.

III. Teager Energy Operator Principles

While virtually all speech processing systems for analysis, synthesis, and recognition employ a linear plane wave model illustrated in Fig. 1(a), it is suggested that under speech pathology conditions where the problem is located at the vocal folds that the nonlinear fluid dynamics interpretation of sound propagation along the vocal tract shown in Fig. 1(b) is more appropriate. It has also been suggested that vortex shedding and flow turbulence are phenomena that should be considered for a more complete understanding of the mechanisms of sound production [8], [36]. Sound could be influenced significantly by these phenomena. The idea of modeling additional sources of acoustic energy may help improve intelligibility and voice quality, while at the same time may provide key elements in voice quality assessment for clinical applications. These ideas on vortex shedding and separated flow were first introduced by Teager and Teager [36]. They performed a series of experiments to measure the dynamic flow field in the vocal tract [36], using hot wire anemometry [35]. Their experimental results strongly suggest nonlinear processes to be the primary sound producing mechanisms in the vocal tract during phonation, and that separated flow and the generated flow vortexes within the confined geometry of the vocal tract are responsible for this phenomena. Later, Thomas, [38] from a computational perspective and McGowan [29] from a theoretical perspective showed, using fluid mechanics and aeroacoustics, that vortices should exist in the vocal tract and that vortices are also a sound source and therefore important in speech production. These studies using fluid mechanics and aeroacoustics focus on the analysis of sound production and propagation in air using equations of air motion which include mass, momentum, and energy conservation. These theories are considered to be more general than source-filter theory because theoretical aeroacoustics is a vector field theory and it accounts for nonlinear excitation sources of sound, whereas the source-filter theory uses scalar field theory and neglects the contribution of the nonlinear excitation sources to the mechanisms of sound production [29].

Speech formants are characterized by the poles of the vocal tract transfer function, when a linear model is used. Teager was convinced that the speech resonances can change rapidly both in frequency and amplitude even within a single pitch period, possibly due to separated airflow in the vocal tract [28].

The nonlinear differential TEO can detect formant AM–FM modulations by estimating the product of their time-varying amplitude and frequency. The Teager operator is considered a high-resolution energy estimator. The AM–FM model proposed by Maragos *et al.* [28] represents a single speech resonance $R(t)$ as an AM–FM signal

$$R(t) = a(t)\cos\left\{2\pi\left[f_c t + \int_0^t q(\tau)\,d\tau\right] + \theta\right\} \qquad (1)$$

where f_c is the formant frequency value, $q(t)$ is the frequency-modulating signal, and $a(t)$ is the time-varying amplitude. The instantaneous formant frequency of the signal is defined as $f_i(t) = f_c + q(t)$. To demodulate a speech resonance $R(t)$ into its varying amplitude $|a(t)|$ and instantaneous frequency $f_i(t)$, the energy separation algorithm (ESA) (developed by Maragos *et al.* [28]) is applied to the signal resonance $R(t)$ obtained after filtering the speech signal around the formant under consideration.

The ESA is based on the Teager energy tracking operator

$$\Psi[s(t)] = \left[\dot{s}(t)^2 - s(t)\ddot{s}(t)\right] \qquad (2)$$

where $\dot{s}(t) = ds/dt$.

When the analogous approach is performed in the discrete time domain, it results in

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1). \qquad (3)$$

At times, it is important to characterize the differential energy from the energy tracking operator in (3). To accomplish this,

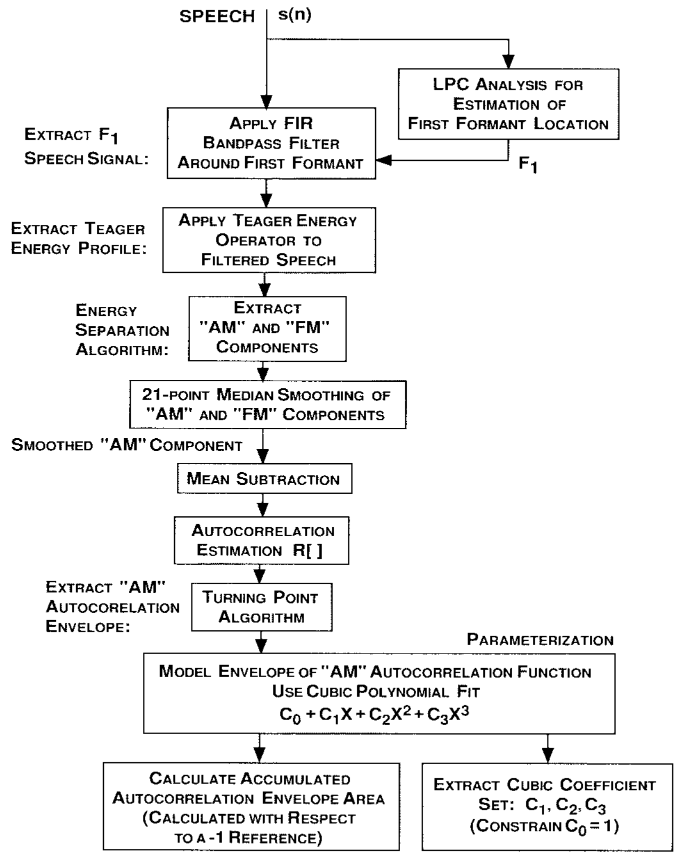

Fig. 2. Flow diagram of the nonlinear speech processing algorithm.

data will be used over a two-sample width (i.e., three data point window) across the input data signal sequence $x(n)$. Maragos *et al.* [28] formulated the ESA frequency and amplitude estimates using the two-sample differential energy separation algorithm (DESA-2) as follows, where the subtraction is over two sample periods[1]:

$$f(n) \approx \frac{1}{4\pi T}\arccos\left\{1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]}\right\} \qquad (4)$$

$$|a(n)| \approx \frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1) - x(n-1)]}}. \qquad (5)$$

It should be noted that errors do occur in the estimate of the FM (4) and AM [from (5)] responses. One of the reasons for this is the lower limit threshold which must be applied to (3). Note that since (3) is the discretized version of (2), differences will exist between the analog and discrete versions $\Psi[s(t)]$, $\Psi[x(n)]$ of the energy profile. Specifically, there is also a small chance that $\Psi[x(n)]$ can be negative if $x(n)$ is a minimum and $x(n+1)$, $x(n-1)$ are both larger. Since $\Psi[x(n)]$ is an energy signal, it must be positive and, therefore, the discrete version of the TEO normally uses a lower threshold. These factors can at times contribute to discontinuities in the AM–FM signals (4) and (5). To address this issue, median filtering is normally applied to reduce the impact of these discontinuities.

[1] Note that in (4), the TEO is applied over a longer time width of two sample periods $x(n+1) - x(n-1)$. This longer time spread is needed to obtain the *differential* TEO and, therefore, the reason for dividing by two.

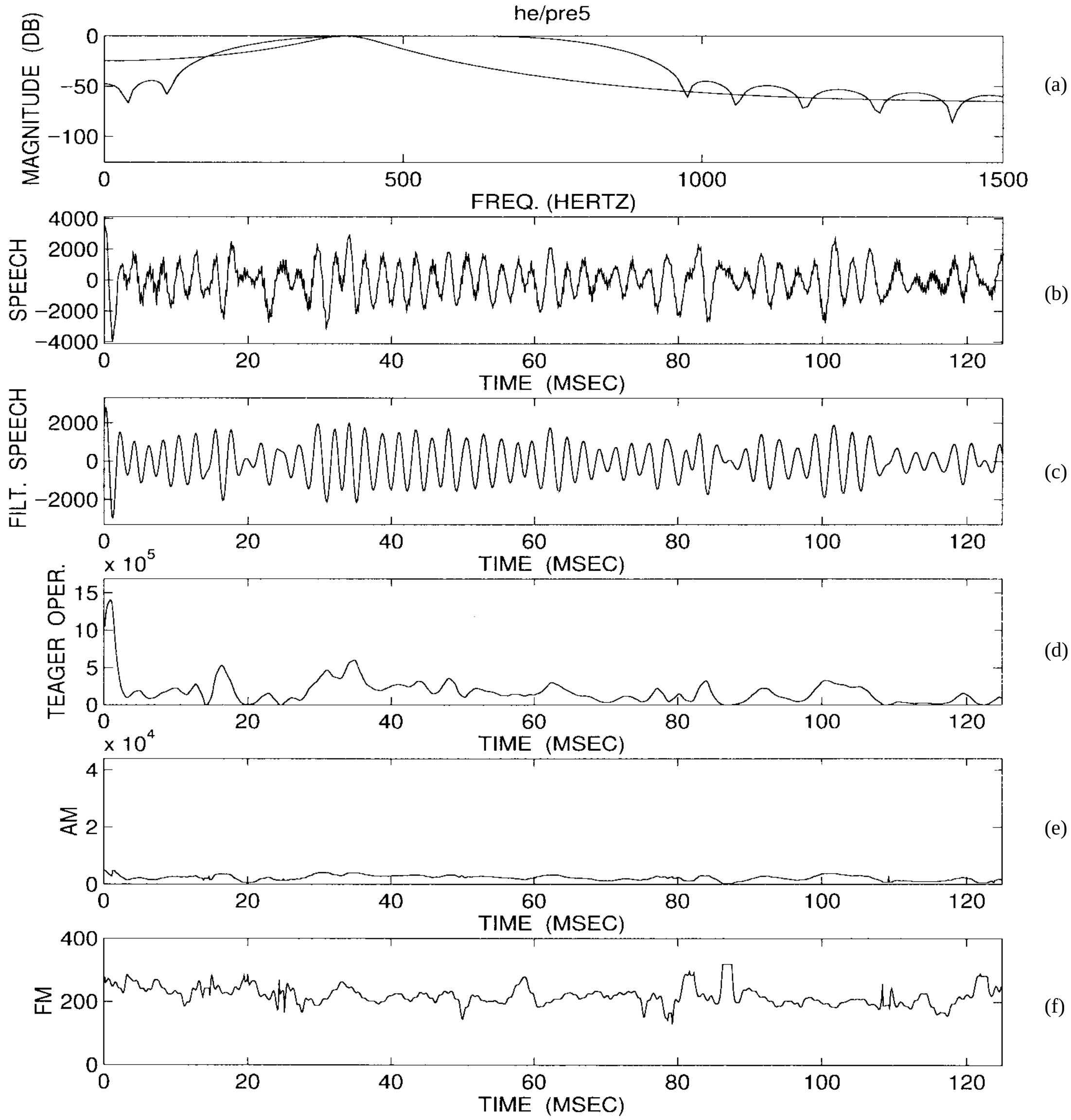

Fig. 3. Nonlinear speech analysis results for a *Female speaker* with muscular tension dysphonia under *Pre*-voice therapy treatment. Shown (from top to bottom) are (a) the frequency content of the speech signal around first formant, and the frequency response of the bandpass filter used. Also shown are the (b) original speech and (c) bandpass filtered speech, (d) TEO profile, and the first formant (e) AM and (f) FM modulations.

IV. Algorithm Formulation

Traditional speech processing methods for laryngeal pathology assessment generally assume linear speech production with measures derived from an estimated glottal flow waveform. They normally require the speaker to achieve complete glottal closure which for many vocal fold pathologies cannot be accomplished. To address this issue, a nonlinear signal processing approach is proposed which does not require direct estimation of the glottal flow waveform. It is noted that glottal flow is not as important here, since it is the effect downstream from the vocal folds which impacts the characteristics of the resulting speech sound. This technique is motivated by earlier studies of airflow characterization for human speech production. The proposed nonlinear approach employs a differential TEO and the energy separation algorithm to obtain formant AM and FM modulations from filtered speech recordings. A new speech measure is proposed based on parameterization of the autocorrelation envelope of the AM response.

A flow diagram of the proposed nonlinear procedure is illustrated in Fig. 2. A tenth-order LPC (linear predictive coefficient) spectrum analysis is performed on the speech signal in order to extract the first formant (F_1) location. The first formant was specifically chosen because this formant is usually associated with the region just above the vocal folds where any effects of vocal fold pathology will first be noticed. A finite impulse response (FIR) bandpass filter is applied to the original speech signal around the first formant of the original speech to isolate the formant. The FIR filter is selected since it results in no phase distortion with frequency. A fixed filter bandwidth of 450 Hz is chosen for all processing. The choice of the Gabor filter bandwidth is typically selected to capture speech energy around the first formant, assuming good separation of the first and second formants. If other formants

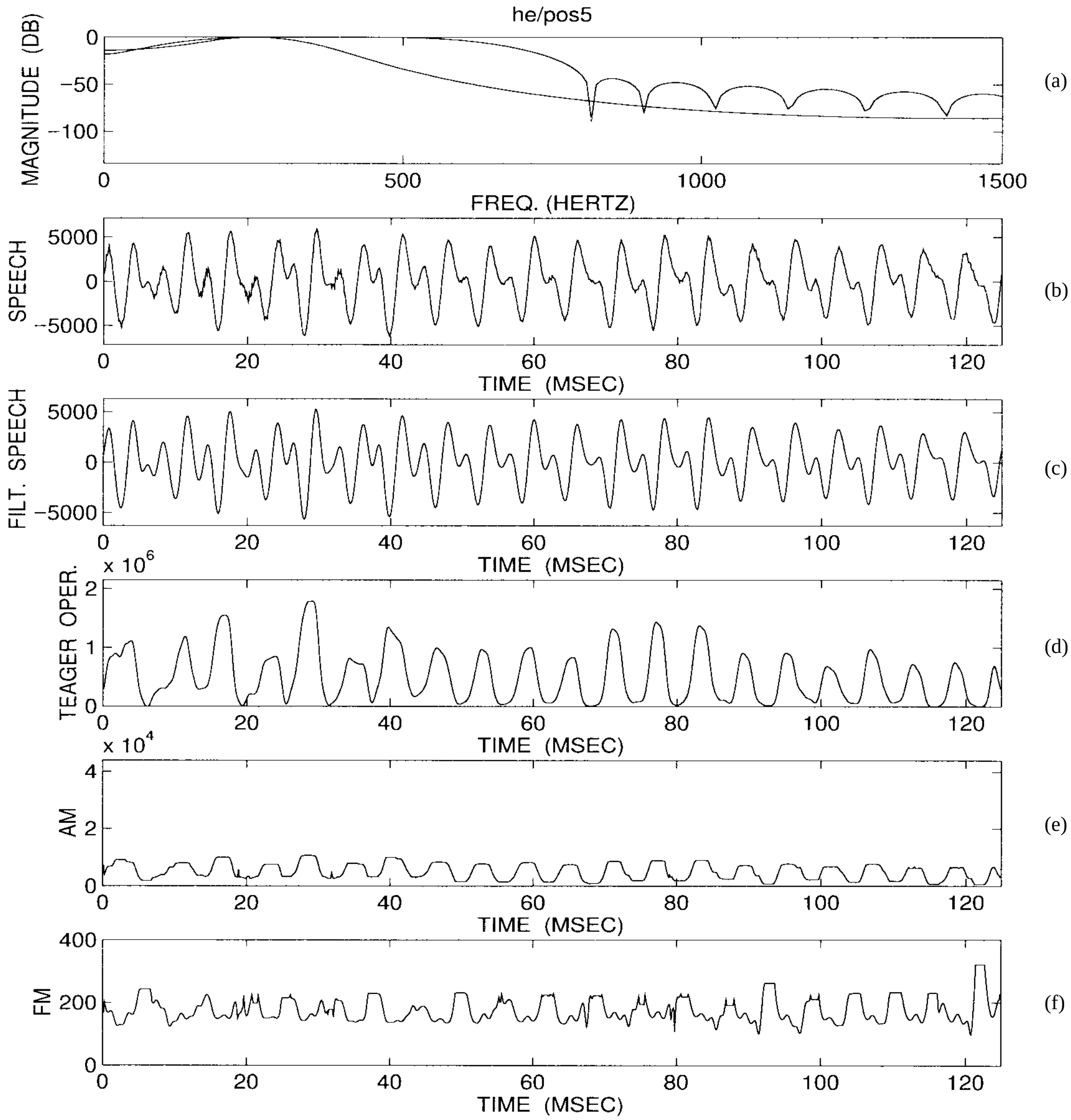

Fig. 4. Nonlinear speech analysis results for a *Female speaker* with muscular tension dysphonia under *Post*-voice therapy treatment. Shown (from top to bottom) are (a) the frequency content of the speech signal around first formant, and the frequency response of the bandpass filter used. Also shown are the (b) original speech and (c) bandpass filtered speech, (d) TEO profile, and the first formant (e) AM and (f) FM modulations.

are considered, or if phonemes which do not contain well separated formants are considered, further studies are needed to select the bandwidth for the bandpass filter.

Next, an estimate of the AM and FM modulation in the first formant time signal is obtained from the TEO of both the filtered speech and its derivative using the energy separation algorithm developed by Kaiser [22] and Maragos *et al*. [28]. A 21-point median smoothing is applied to both the AM and FM component profiles, at those locations where the value of the computed modulated signals exceed the overall median value by 40%. This smoothing is needed, since it is possible to obtain discontinuities in the AM or FM responses as discussed in the previous section.

Pitch information is obtained from the AM signal by determining the location where there is a subsequent maximum in the AM autocorrelation function (i.e., the first maximum after the zeroth lag point $k = 0$). A window of three pitch periods is used to calculate the mean of the smoothed AM component, which is then subtracted from the input smoothed AM response.

Next, the autocorrelation function of the AM envelope is obtained. If a sequence $s(n)$ is of length N, the resulting linear autocorrelation has length $N-1$, and can be obtained as follows:

$$R_{ss}(k) = \mathcal{FT}^{-1}\left[|S'(k)|^2\right] \tag{6}$$

where $S'(k)$ represents the Fourier transform of the zero-padded version of the sequence $s(n)$, and $R_{ss}(k)$ represents the linear autocorrelation of the sequence $s(n)$.

To obtain an estimate of the envelope of the AM component, a peak-picking turning point algorithm [40] is employed. It was determined from early analysis of normal and vocal fold pathology speech data that strong periodicity remained in the autocorrelation function of the smoothed AM response, but

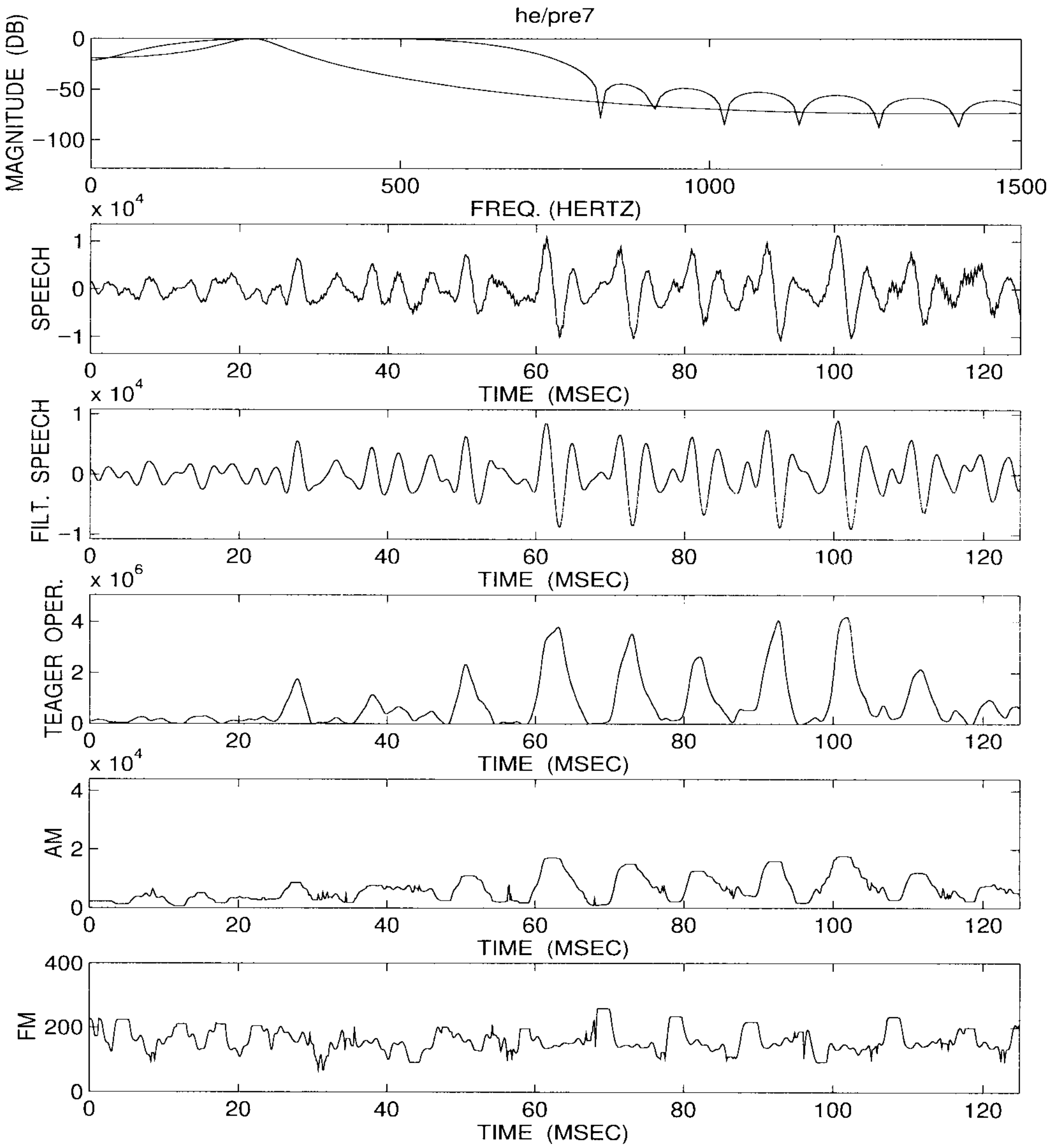

Fig. 5. Nonlinear speech analysis results for a *Male speaker* with muscular tension dysphonia under *Pre*-voice therapy treatment. Shown (from top to bottom) are (a) the frequency content of the speech signal around first formant, and the frequency response of the bandpass filter used. Also shown are the (b) original speech and (c) bandpass filtered speech, (d) TEO profile, and the first formant (e) AM and (f) FM modulations.

that the degree of correlation dropped significantly with a loss of periodicity for vocal fold pathology speech. A simple slope calculation is performed to determine the instants where changes in the sign of the slope occur, which identify the peaks (turning points) of the signal.

Once the envelope of the autocorrelation function of the smoothed AM Teager profile is obtained, the response is parameterized for analysis between healthy and vocal fold pathology speakers. Two methods are considered, a cubic polynomial fit and an accumulated envelope area measure. In the first approach, the envelope of the AM component is modeled using the following cubic fit polynomial

$$p(X) = C_0 + C_1 X + C_2 X^2 + C_3 X^3. \tag{7}$$

The coefficients C_1, C_2, and C_3 are obtained under healthy and pathology conditions. Here, we constrain C_0 to be one, since for a valid autocorrelation function the cubic polynomial model must be unity for $X = 0$. For the second approach, the equation for the polynomial fit under healthy and pathology conditions was used to calculate the accumulated autocorrelation envelope area. The area was calculated with respect to a -1 reference value.

V. Algorithm Evaluation

A. Laryngeal Pathology Speech Data

The motivation here for algorithm evaluation is to develop an understanding of how appropriate the proposed nonlinear speech processing algorithm would be in a clinical setting. The evaluations here are not intended to confirm the reliability of the method in clinical speech pathology assessment, but merely to probe the usefulness of the method for a range of vocal fold speech pathologies. More extensive evaluations would be

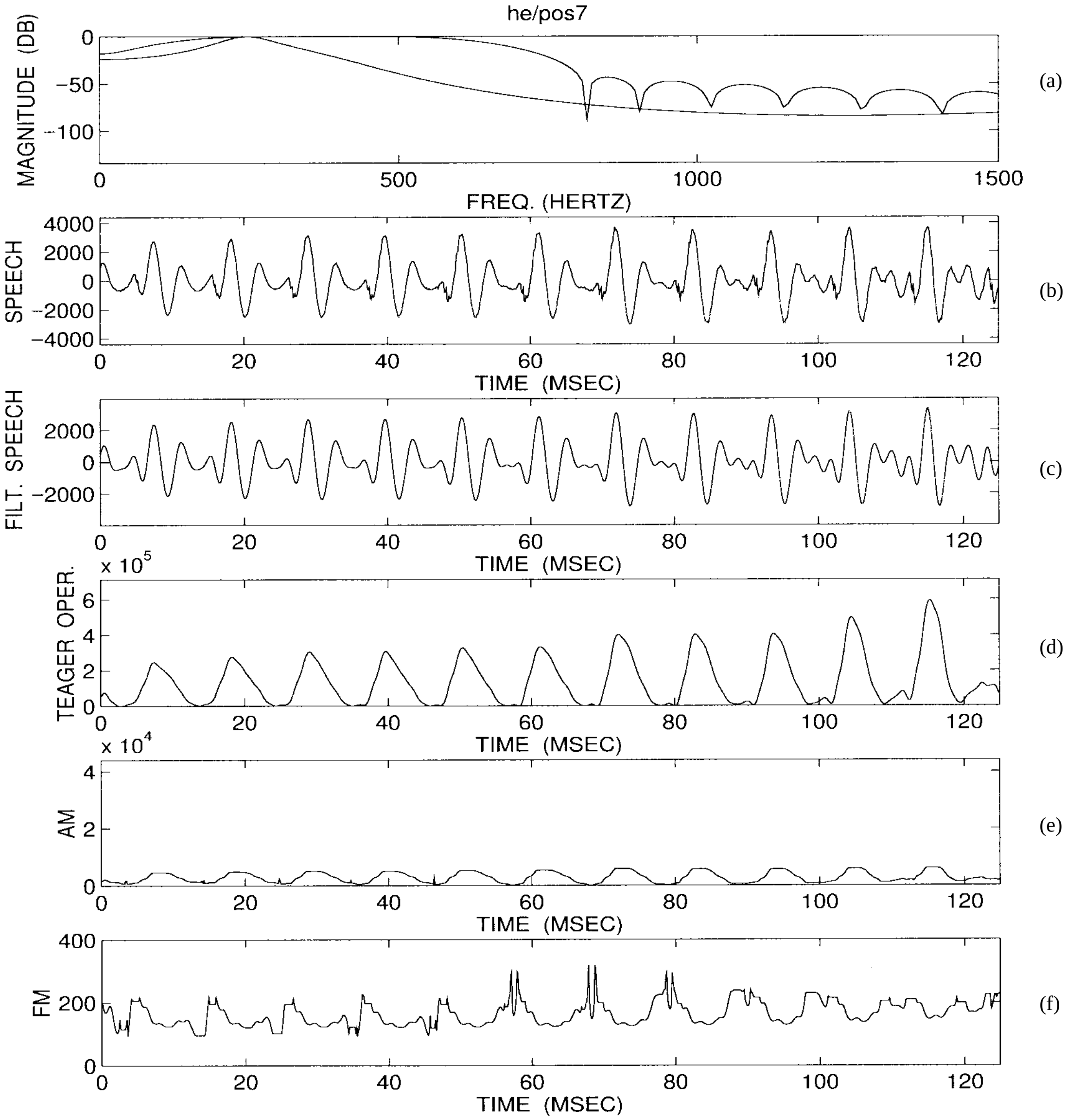

Fig. 6. Nonlinear speech analysis results for a *Male speaker* with muscular tension dysphonia under *Post*-voice therapy treatment. Shown (from top to bottom) are (a) the frequency content of the speech signal around first formant, and the frequency response of the bandpass filter used. Also shown are the (b) original speech and (c) bandpass filtered speech, (d) TEO profile, and the first formant (e) AM and (f) FM modulations.

needed to extend the proposed algorithm as a clinical speech pathology tool.

A corpus of speech data from adult speakers was selected by an experienced clinical speech pathologist from a speech pathology library of pre- and post- therapy vocal sessions. The examples represent a broad range of speech quality as judged by the speech pathologist in pre-therapy conditions, and is therefore intended to probe the ability of the nonlinear speech processing algorithm for a broad range of vocal fold pathologies. The test corpus consisted of 11 adult speakers (ten female, one male) producing the vowel /e/ in the phrase "he is," extracted from the *Grandfather passage*, a text fragment commonly used by speech therapists to assess patient voice quality. The front vowel /e/ was specifically chosen because it possesses first and second formants that are widely separated (over 1000 Hz separation). Therefore, the AM–FM modulations of the second formant will not measurably influence the first formant AM–FM modulations measured with the proposed nonlinear speech processing technique. From the 11 speakers, ten suffered from functional voice disorders, or more specifically muscular tension dysphonias (MTD), and one was completely healthy. It was noted that some speakers had recurrent laryngitis as a result of vocal abuse. Specific characteristics of these MTD patients included 1) sudden onset of "hysterical" vocal hyperfunction, 2) voiceless and upward pitch breaks, 3) recurring hoarseness, 4) laryngitis, and 5) dysphonic to aphonic voice. For this speaker set, repeated tokens of the /e/ phoneme were extracted during the initial patient evaluation. Next, a period of voice therapy was performed by a speech pathologist (normally 15–45 min), after which a second recording of the "Grandfather passage" was obtained. The speech pathologist noted that significant improvement in voice quality resulted from voice therapy in nine of the 11 patients (the two remaining patients also

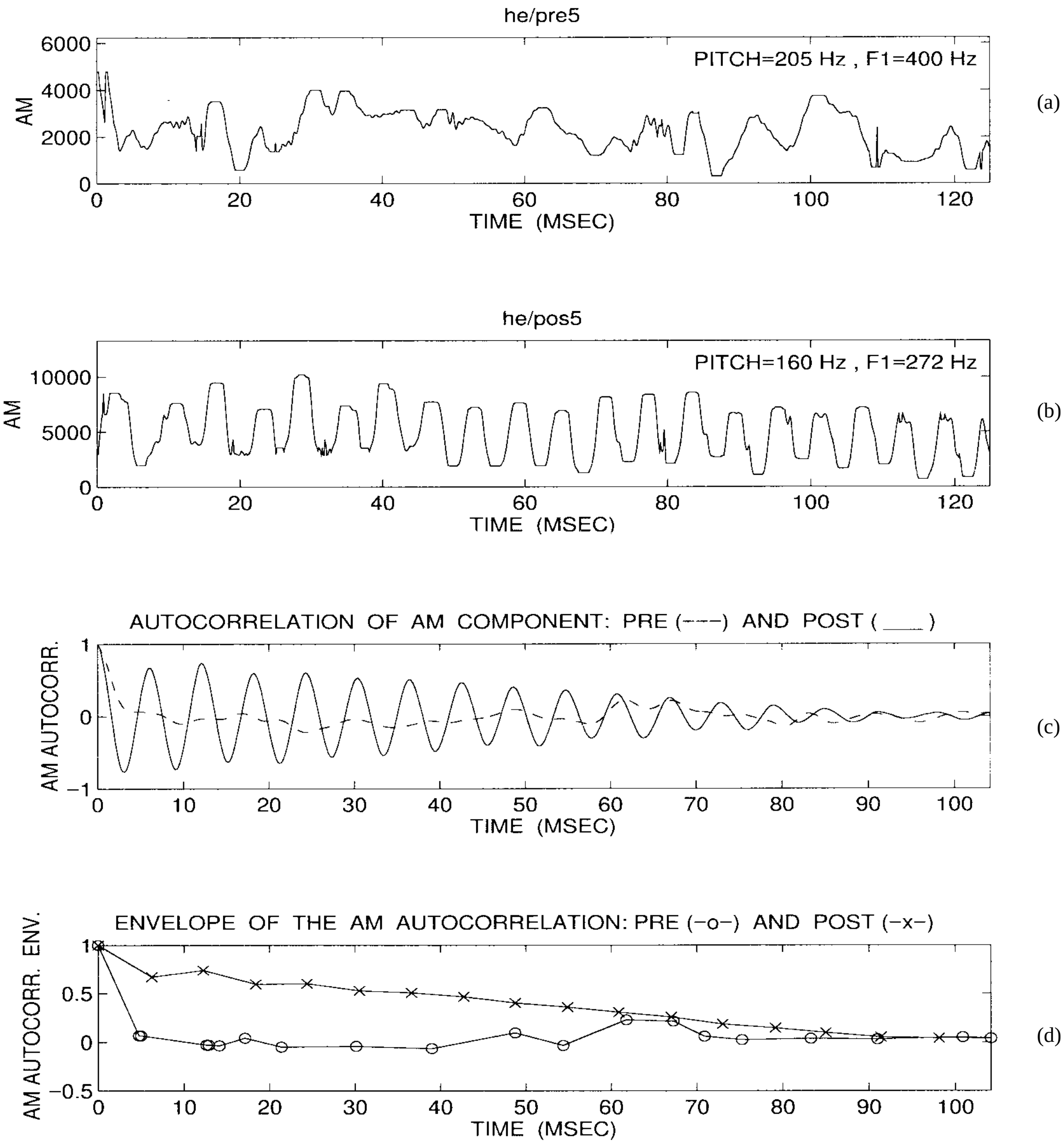

Fig. 7. First formant AM modulation analysis of (a) pre- and (b) post-voice therapy treatment for a *female* speaker with muscular tension dysphonia. Also shown are the (c) pre- and post-AM autocorrelation functions, and (d) AM autocorrelation envelopes. Pre- and post-AM autocorrelation peaks are marked with "o" and "x," respectively.

showed improvement but their resulting speech was judged to be slightly breathy). All speech data was digitized at an 8-kHz sampling rate, using a 16-bit A/D converter.

B. Feature Profile Results

The nonlinear-based speech analysis algorithm was applied to all speakers in the database under both pre- and post-therapy conditions. For the purpose of demonstrating detailed algorithm performance, we consider specific results for two of the speakers (one male, one female). Figs. 3–6 illustrate the results obtained for these two patients with MTD pre- and post-voice therapy treatment.

In Figs. 3 and 5, pre-voice therapy results for a female and male speaker are shown. Plot (a) shows the frequency content of the speech signal around the first formant, and the frequency response of the bandpass filter used to extract this speech portion information. In plot (b), a portion of the original speech signal for the /e/ phoneme is shown, followed by (c) the extracted speech signal after application of the bandpass filter centered around F_1 for this speaker. In plot (d), we show the profile obtained after using the TEO on the bandpass filtered speech. The extracted AM and FM components from the TEO are shown in plots (e) and (f), respectively. Figs. 4 and 6 summarize equivalent data responses for the same female and male speaker after post-voice therapy. After evaluating each of the speakers in the database under pre- and post-voice therapy conditions, it was determined that the AM component was better able to convey the presence of consistent periodic structure, believed to be associated with the quality of speech under a muscular tension dysphonia. This is demonstrated in the two sample patients if we compare Figs. 3(e) with 4(e), and 5(e) with 6(e). When more normal, regular speech production occurs in the original speech waveform, the corresponding AM response is more regular, with less noiselike or random

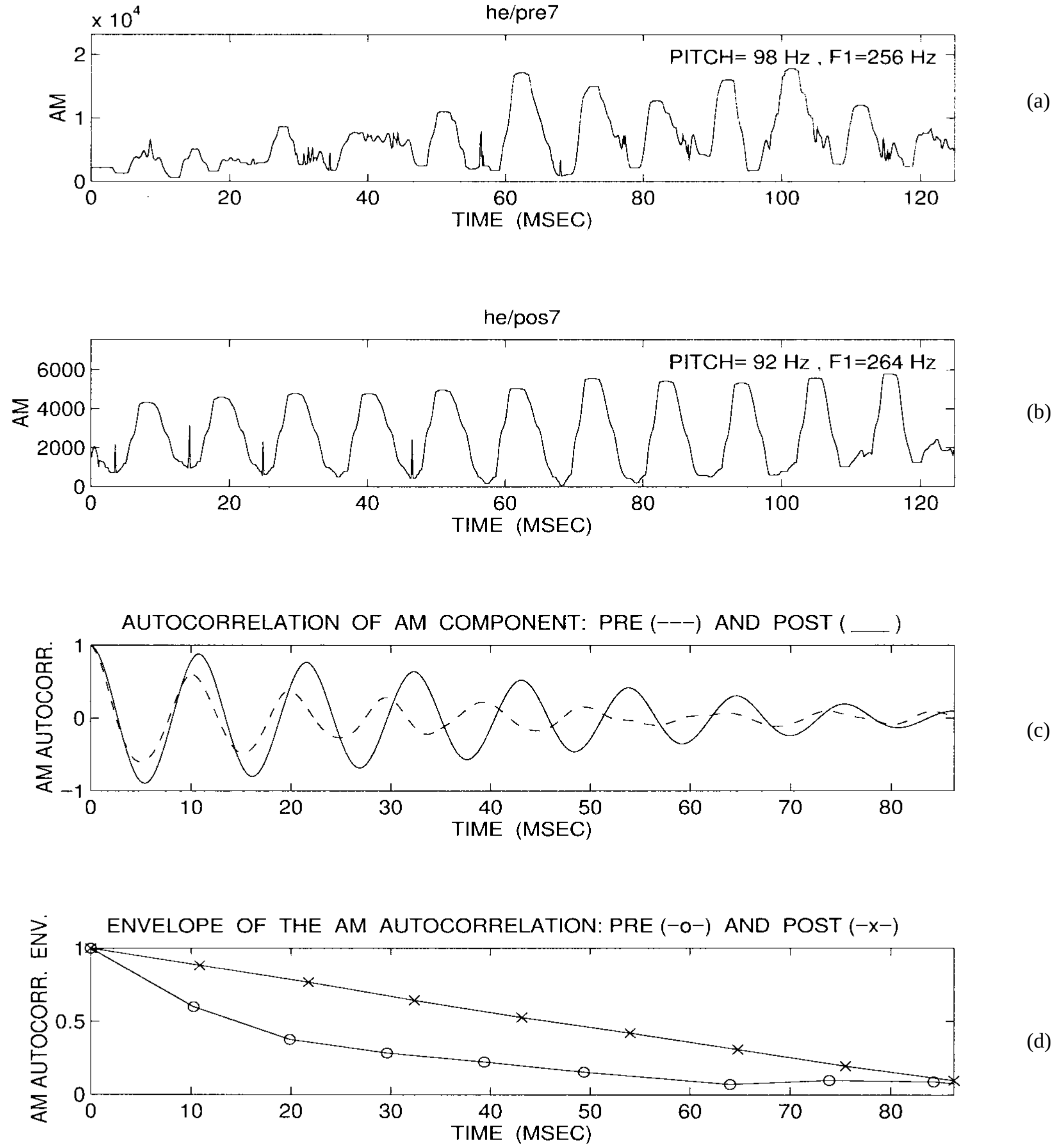

Fig. 8. First formant AM modulation (a) pre- and (b) post-voice therapy treatment for a *male* speaker with muscular tension dysphonia. Also shown are the (c) pre- and post-AM autocorrelation function and (d) AM autocorrelation envelopes. Pre- and post-AM autocorrelation peaks are marked with "o" and "x," respectively.

behavior. It is suggested that this marked change in the periodicity and regularity of the AM component after voice therapy, is an indication of successful treatment.

C. Quantitative Feature Assessment

In order to quantify this distinctive feature, the autocorrelation function of the AM component was obtained. It is suggested that the decay in the successive peak values of the AM autocorrelation function represents a measure of the regularity and periodicity of the signal. There is an inherent decay that arises from the fact that the autocorrelation estimate is the biased estimate, so the total number of points used to calculate the autocorrelation decreases for each successive lag. However, this fact does not affect the performance of the processing technique, since its impact on both pre- and post-therapy signals is the same. If an unbiased estimate is used instead, the envelope of the AM autocorrelation will be more flat, but the separation between both pre- and post-voice therapy treatment groups is still present. The biased autocorrelation estimate has the added advantage that it can be obtained using the fast Fourier transform, which saves computational time.

To illustrate quantitative feature assessment, we consider a sample female and male speaker. In Fig. 7, we summarize the resulting AM responses for (a) pre- and (b) post-voice therapy treatment for a female speaker [corresponding responses for a male speaker are shown in Fig. 8(a) and (b), respectively]. Again, the clear differences in nonlinear excitation structure is apparent. Next, we plot the mean normalized AM autocorrelation function for both pre- and post-therapy conditions [Female Fig. 7(c), male Fig. 8(c)]. Using the turning point algorithm, the local maxima are extracted from plot (c), and a piecewise linear approximation of the AM autocorrelation

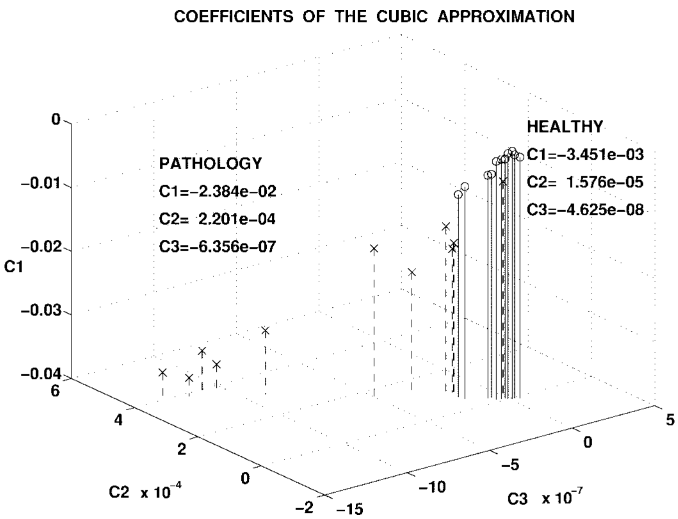

Fig. 9. Scatter plots for healthy and pathology conditions in the parametric C_1, C_2, C_3 space.

TABLE I

Parameterized Models for AM Autocorrelation Responses

C_1 $(\times 10^{-3})$		C_2 $(\times 10^{-5})$		C_3 $(\times 10^{-7})$	
Pre	Post	Pre	Post	Pre	Post
−17.40	−2.582	9.250	1.260	− 1.650	−0.530
−20.95	−4.970	13.71	2.753	− 3.249	−0.757
−29.92	−6.588	30.27	5.056	− 8.933	−1.700
−13.58	−4.805	9.225	2.376	− 2.047	−0.605
−37.09	−8.133	38.54	6.622	−11.97	−1.799
−35.28	−1.701	36.17	−0.513	− 1.077	0.241
− 6.088	−1.421	1.469	0.109	− 0.075	−0.036
−32.64	−2.337	36.42	0.535	−11.59	−0.137
−16.69	−1.049	16.78	−0.480	− 4.966	0.105
−16.47	−2.336	9.068	0.841	− 1.602	−0.302
−36.12	−2.039	41.22	−1.227	−13.06	0.432
Coefficient Means					
$m_{C_1}(\times 10^{-3})$		$m_{C_2}(\times 10^{-5})$		$m_{C_3}(\times 10^{-7})$	
−23.84	−3.451	22.01	1.576	−6.356	−0.465
Coefficient Standard Deviations					
$\sigma_{C_1}(\times 10^{-3})$		$\sigma_{C_2}(\times 10^{-5})$		$\sigma_{C_3}(\times 10^{-7})$	
10.71	2.326	14.60	2.448	4.943	0.732

envelope is plotted in (d). We note that for both speakers, as well as for the others in the database, the AM autocorrelation envelope for all post-voice therapy cases can be approximated by a straight line, and its amplitude is noticeably higher than those corresponding to the pathology case (pretreatment). This is clearly due to change in the excitation (both linear and suggested nonlinear) characteristics, since the vocal tract response under both conditions are from the same speaker. In effect, the differences in the decay of the AM envelope response is reflecting a change in the regularity of the energy response. We also point out that for the pathology case, the AM autocorrelation response decreases significantly and often exhibits irregularities after 5–40 ms, depending on the level of pathology (muscular tension dysphonia) present in the speaker's voice under pre-voice therapy conditions.

Next, the envelope of the AM component was modeled using the cubic polynomial $p(X) = C_0 + C_1X + C_2X^2 + C_3X^3$. The second and third-order coefficients C_2 and C_3 are expected to be very small for the post-therapy cases, since the shape follows primarily a linear trend. The coefficient C_0 was set to one since the value of the autocorrelation for lag zero is constrained to be one.

Table I shows the set of coefficients C_1, C_2, and C_3 obtained for the cubic fit of the autocorrelation envelope for the 11 patients considered in this study. Note the smaller order of magnitude of the coefficients for the post (healthy) cases. A linear approximation discarding C_2 and C_3 in the post cases is still adequate to represent the envelope. This is not the case for the pre (pathology) cases. The means and standard deviations of each model parameter are shown for pre- and post-therapy conditions. Since the range of speech quality from MTD pathologies was large, the variation in model responses will be larger (i.e., parameter standard deviations are large in the pre-therapy conditions). After therapy, all subjects showed marked increases in speech quality, which is confirmed in the significant reduction in model parameter variation.

By employing results from the cubic polynomial fit, we can assess quantitatively how the shape of the AM autocorrelation envelope is effected between healthy and pathology conditions. Fig. 9 shows a spatial distribution on a three-dimensional plane of the pre- and post-therapy cases. The coefficients C_1, C_2, and C_3 are represented as the projection on the x, y, and z axes, respectively. All healthy cases tend to group in a tight cluster, whereas the pathology cases are scattered throughout the parameter space. Observations from a clinical assessment of speech pathology for these muscular tension dysphonia patients indicates that they possessed a wide range of voice quality (i.e., varying levels of hoarseness, breathiness, etc.) while more consistent voice quality resulted after voice therapy.[2] This helps explain the reason for a wider feature variation in pre- versus post-speech conditions. This figure, therefore, illustrates how healthy and nonhealthy speech cases are distributed in the C_1, C_2, C_3 space. We point out however that without independent subjective assessment methods, it is not possible at this time to quantify how overall speech quality is improved as the response moves from pathology locations (i.e., marked with "x"), toward healthy (marked with "o").

Fig. 10 shows the AM autocorrelation envelopes of all pre- and post-cases using a cubic fit polynomial. From the figure, it is evident that to be able to model the entire 25-ms segment using a single polynomial fit in each case, a higher-order model is required for the pathology case, whereas a linear fit is adequate to model the healthy cases. Again, note the clear separation between healthy and pathology groups. In particular, we point out that the initial slope in the autocorrelation response for the first 5–10 ms would in general be sufficient to distinguish between healthy and severe pathology cases. This is confirmed by the large negative values for coefficient C_1 (i.e., slope) in Table I in the pre-therapy condition, and the reduced values for post-therapy condition. Fig. 11 shows the points used to estimate the overall model of the AM autocorrelation envelope for both healthy and pathology cases. The average cubic fit

[2] It is interesting to note that the two patients who responded to a lessor degree after voice therapy are somewhat closer to those pathology cases in Fig. 9.

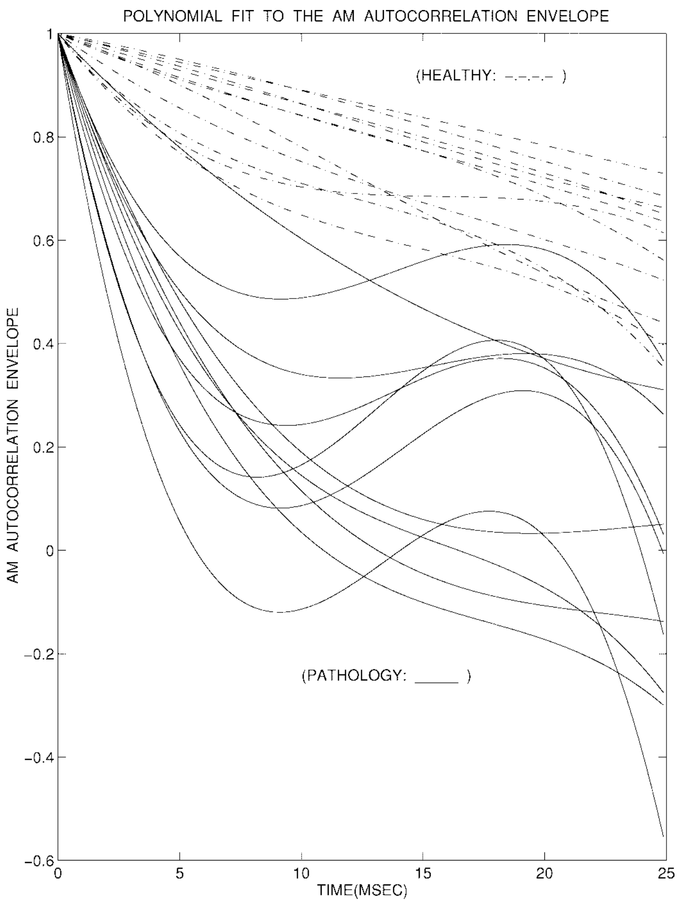

Fig. 10. Cubic polynomial fits from the first formant AM modulation for healthy and pathology conditions (all 11 test subjects are shown).

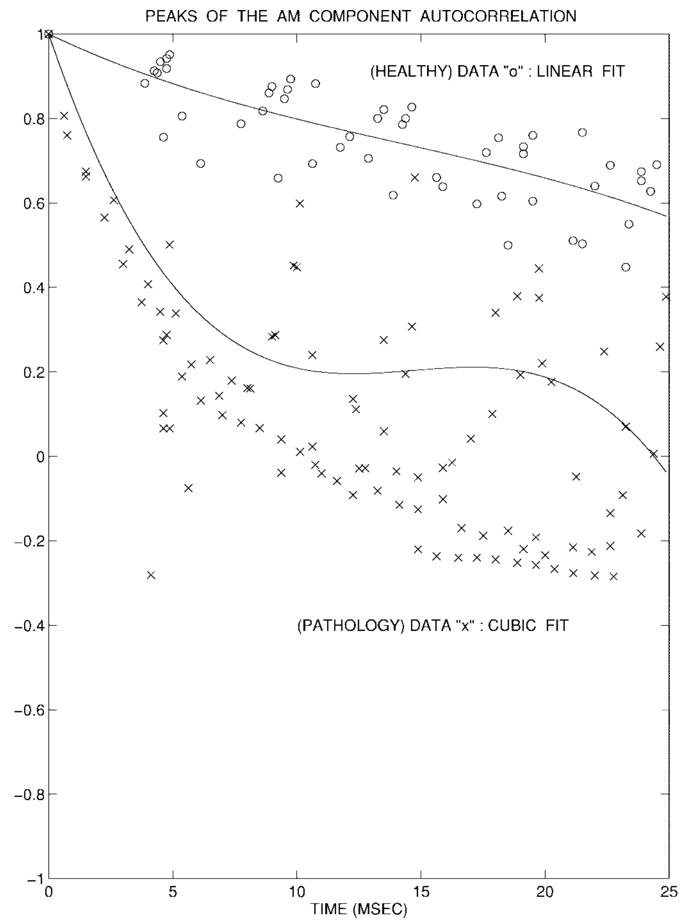

Fig. 11. Local maxima in the autocorrelation envelope of first formant AM modulation for healthy and pathology conditions on the 11 subjects used in this study. Also shown are the autocorrelation envelopes for healthy and pathology conditions obtained using a model with the averaged C_1, C_2, and C_3 coefficients for each condition, respectively.

obtained for each condition represents a general model in which to compare the variation in healthy versus vocal fold pathology conditions. Again, in the healthy case, the model can be reduced to a linear fit.

In our final assessment, we obtain an overall area measure of the envelope of the AM autocorrelation response. The AM autocorrelation response was interpolated using the corresponding cubic fit in each case, and the result was used to obtain the accumulated area. This area was calculated with respect to the -1 reference for the first 25 ms. Fig. 12 shows the results obtained for all speakers in the database under pre- and post-therapy conditions. In this evaluation, perfect healthy/pathology classification is obtained with an area threshold of 325. It is clear that the area parameter can, therefore, be used to assess or classify speakers under healthy or pathology due to functional voice disorders.

VI. Conclusions

A nonlinear speech processing technique for extraction of speech parameters has been proposed and developed in the analysis of speech with MTD. Evaluation of this algorithm for both pre- and post-voice therapy treatment showed that the analysis of the first formant AM modulation characteristic allows the extraction of features that we believe to be correlated to the regularity of vocal fold vibratory movement in a healthy condition and to the asymmetry and irregular structure for a vocal fold pathology condition. The total number of multiplications required with this procedure is $N(\log N+8)$, and N square roots for N samples. Therefore, this procedure could be suitable for real-time applications due to its computational simplicity. Two measures were proposed, one based on the cubic parameter values, and one based on the accumulated area from the autocorrelation of AM response. Both clearly demonstrated the effectiveness of the nonlinear approach. The proposed method could also be extended for larger speaker populations, for speech pathology assessment based on age, gender, and phoneme. In addition, further research is suggested to develop objective speech quality measures [33] based on the degree of parameter variation using features obtained with this procedure. An example would be an in-depth study assessing the degree of quality variation with parameter movement from healthy to pathology in the C_1, C_2, C_3 feature space. Clearly, the evaluations presented here can only be used to suggest that the proposed nonlinear processing technique could be effective for vocal fold speech pathology assessment, and that more extensive evaluations are needed to verify if the method would be effective as a clinical assessment tool for speech pathologists.

In closing, this study has resulted in the formulation of a new method which shows that a fast, effective digital speech processing technique can be developed for vocal fold

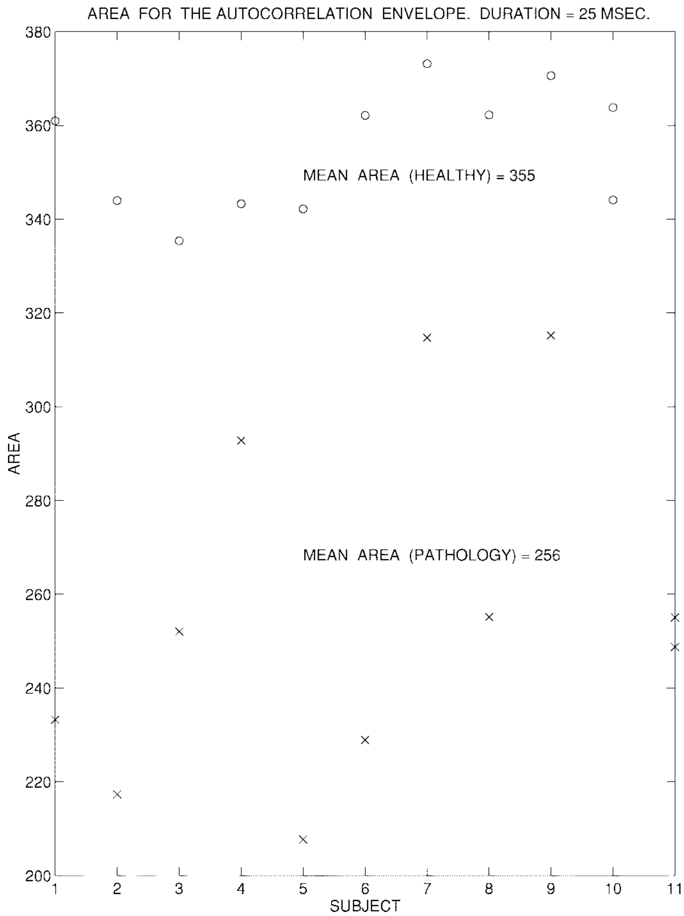

Fig. 12. Accumulated area for the autocorrelation envelope of first formant AM modulations for the 11 speakers used in this study.

pathology assessment, without the need for direct glottal flow estimation. This technique is both 1) noninvasive and 2) does not require estimating the instant of glottal closure. This study also suggests that alternative nonlinear methods can be proposed which begin to address the limitations of previous linear approaches to speech pathology assessment.

Acknowledgment

The authors would like to thank Dr. J. E. Riski, Director of the Speech Pathology Laboratory, Center for Craniofacial Disorders, Scottish Rite Children's Medical Center, Atlanta, GA, for making vocal fold pathology speech data available for use in this study.

References

[1] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, nos. 2/3, pp. 109–118, June 1992.

[2] A. E. Aronson, "Motor speech signs of neurologic disease," in *Speech Evaluation in Medicine*, J. Darby, Ed. New York: Grune and Stratton, 1981.

[3] D. M. Bless, "Measurement of vocal function," in *Otolaryngologic Clinics of North America. Voice Disorders*, J. Koufman and G. Isaacson, Eds. Philadelphia, PA: Saunders, Oct. 1991, vol. 24, no. 5, pp. 1023–1033.

[4] D. A. Cairns, J. H. L. Hansen, and J. E. Riski, "A noninvasive technique for detecting hypernasal-speech using a nonlinear operator," *IEEE Trans. Biomed. Eng.*, vol. 43, pp. 35–45, Jan. 1996.

[5] D. G. Childers, J. M. Naik, J. N. Larar, A. K. Krishnamurthy, and G. P. Moore, "Electroglottography, speech, and ultra-high speed cinematography," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, I. Titze and R. Scherer, Eds. Iowa City, IA: Iowa Univ. Press, 1983, pp. 201–220.

[6] R. Daniloff, G. Schuckers, and L. Feth, *The Physiology of Speech and Hearing. An Introduction.* Englewood Cliffs, NJ: Prentice-Hall, 1980.

[7] J. K. Darby, *Speech Evaluation in Medicine.* New York,: Grune and Stratton, 1981.

[8] P. O. Davies, R. S. McGowan, and C. H. Shadle, "Practical flow duct acoustics applied to the vocal tract," in *Vocal Fold Physiology: Frontiers in Basic Science*, I. R. Titze and R. C. Scherer, Eds. San Diego, CA: Singular, 1993, pp. 93–142.

[9] P. J. Davis, D. Bartlett, and E. S. Luschei, "Coordination of the respiratory and laryngeal systems in breathing and vocalization," in *Vocal Fold Physiology: Frontiers in Basic Science*, I. Titze and R. Scherer, Eds. San Diego, CA: Singular , 1993, pp. 189–226.

[10] S. B. Davis, "Acoustic characteristics of laryngeal pathology," in *Speech Evaluation in Medicine*, J. Darby, Ed. New York: Grune and Stratton, 1981, pp. 77–104.

[11] ——, "Acoustic characteristics of normal and pathological voices," in *Speech & Language. Advances in Basic Research & Practice*, N. Lass, Ed. New York: Academic, 1979, vol. 1.

[12] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Series. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[13] S. Feijoo and C. Hernandez, "Short-term stability measures for the evaluation of vocal quality," *J. Speech, Hearing Res.*, vol. 33, pp. 324–334, June 1990.

[14] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 3rd ed. New York: Springer-Verlag, 1972.

[15] C. E. Gelfer, K. S. Harris, R. Collier, and T. Baer, "Is declination actively controlled?," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, I. Titze and R. Scherer, Eds. Iowa City, IA: Iowa Univ. Press, 1983, pp. 91–104.

[16] W. Hess, *Pitch Determination of Speech Signals. Algorithms and Devices.* Berlin, Germany: Springer-Verlag, 1983.

[17] S. Hiki, S. Imaizumi, M. Hirano, H. Matsushita, and Y. Kakita, "Acoustical analysis for voice disorders," in *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1976, pp. 613–616.

[18] M. Hirano, "The laryngeal examination," in *Speech Evaluation in Medicine*, J. Darby, Ed. New York: Grune and Stratton, 1981.

[19] M. Hirano, S. Hibi, T. Yoshida, H. Kasuya, and Y. Kikuchi, "Acoustic analysis of pathological voice," *Acta Otolaryngologica*, vol. 105, pp. 432–438, 1988.

[20] D. J. Hoyt, J. W. Lettinga, K. A. Leopold, and S. R. Fisher, "The effect of head and neck radiation therapy on voice quality," *Laryngoscope*, vol. 102, pp. 477–480, May 1992.

[21] J. F. Kaiser, "Some observations on vocal tract operation from a fluid flow point of view," in *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, I. Titze and R. Scherer, Eds. Iowa City, IA: Iowa Univ. Press, 1983, pp. 358–386.

[22] ——, "On a simple algorithm to calculate the 'energy' of a signal," in *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1990, pp. 381–384.

[23] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1329–1334, 1986.

[24] P. Kitzing, "Simultaneous photo- and electroglottographic measurements of voice strain," in *Vocal Fold Physiology: Biomech., Acoust. & Phonatory Control*, I. Titze and R. Scherer, Eds. Iowa City, IA: Iowa Univ. Press, 1983, pp. 221–229.

[25] J. A. Koufman, "Approach to the patient with a voice disorder," in *Otolaryngologic Clinics of North America. Voice Disorders.* Philadelphia, PA: Saunders, Oct. 1991, vol. 24, no. 5, pp. 989–998.

[26] J. A. Koufman and P. D. Blalock, "Classification and approach to patients with functional voice disorders," *Ann. Otology, Rhinol, Laryngol.*, vol. 91, no. 4, pp. 372–377, July 1982.

[27] ——, "Functional voice disorders," in *Otolaryngologic Clinics of North America. Voice Disorders*, Philadelphia, PA: Saunders, Oct. 1991, vol. 24, no. 5, pp. 1059–1073.

[28] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[29] R. S. McGowan, "An aeroacoustic approach to phonation," *J. Acoust. Soc. Amer.*, vol. 83, pp. 696–704, 1988.

[30] T. Murry and E. T. Doherty, "Selected acoustic characteristics of pathological and normal speakers," *J. Speech Hearing Res.*, vol. 23, pp. 361–369, 1980.

[31] D. O'Shaughnessy, *Speech Communication: Human and Machine.* Reading, MA: Addison-Wesley, 1987.

[32] R. J. Prater, "Voice therapy techniques and applications," in *Otolaryngologic Clinics of North America. Voice Disorders,* J. Koufman and G. Isaacson, Eds. Philadelphia, PA: Saunders, Oct. 1991, vol. 24, no. 5, pp. 1075–1092.
[33] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
[34] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice Hall, 1978.
[35] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust. Speech, Signal Processing,* vol. ASSP-28, no. 5, pp. 599–601, Oct. 1980.
[36] H. M. Teager and S. M. Teager, "A phenomenological model for vowel production in the vocal tract," in *Speech Sciences: Recent Advances,* R. G. Daniloff, Ed. San Diego, CA: College-Hill, 1985, pp. 73–109.
[37] ——, "Evidence for nonlinear production mechanisms in the vocal tract," in *Speech Production and Modeling,* Chateau Bonas, France, NATO Advanced Study Institute, July 17–29, 1989.
[38] T. J. Thomas, "A finite element model of fluid flow in the vocal tract," *Comput., Speech, Language,* vol. 1, pp. 131–151, 1986.
[39] I. Titze, *Principles of Voice Production.* Englewood Cliffs, NJ: Prentice-Hall, 1994.
[40] W. J. Tompkins and J. G. Webster, Eds. *Design of Microcomputer-Based Medical Instrumentation.* Englewood Cliffs, NJ: Prentice-Hall, 1981.
[41] D. E. Veeneman and S. L. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 369–377, Apr. 1985.
[42] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-27, pp. 350–355, Aug. 1979.
[43] W. R. Zemlin, *Speech and Hearing Science. Anatomy and Physiology,* 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1988.

John H. L. Hansen (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ, in 1982. He received the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

In 1988, he joined the faculty of Duke University, Durham, NC, the Department of Electrical Engineering, where he established and directs the Robust Speech Processing Laboratory. He has since received a secondary appointment in the Department of Biomedical Engineering. Prior to joining the Duke faculty, he was employed by the RCA Solid State Division, Somerville, NJ (1981 to 1982), and Dranetz Engineering Laboratories, Edison, NJ (1978 to 1981). He has served as a Technical Consultant to industry and the U.S. Government, including AT&T Bell Laboratories, IBM, VeriVoice, DOD, and Sparta in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech analysis. His research interests span the areas of digital signal processing, analysis and modeling of speech under stress or pathology, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust speech enhancement for voice communications, and source generator-based speech modeling for robust recognition in stress, noise, and Lombard effect. He is the author of numerous journal papers in the field of speech processing and communications and is co-author of the textbook *Discrete-Time Processing Speech Signals,* (Prentice-Hall, 1993).

Dr. Hansen was the recipient of a Whitaker Foundation Biomedical Research Award in 1993, a National Science Foundation's Research Initiation Award in 1990, and has been named a Lilly Foundation Teaching Fellow. He was an invited tutorial speaker for ICASSP-95 and the ESCA-NATO Speech Under Stress Research Workshop, Lisbon, Portugal. He has served as Chairman for the IEEE Communications and Signal Processing Society of North Carolina (1992 to 1994), Advisor for the Duke University IEEE Student Branch (1990 to 1997), Tutorials Chair for ICASSP-96: *International Conference on Acoustics, Speech, and Signal Processing,* and is presently serving as Associated Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992 to 1997). He has also served as Guest Editor of the Oct. 1994 special issue on Robust Speech Recognition for that publication.

Liliana Gavidia-Ceballos was born in Caracas, Venezuela, on June 6, 1965. She received the B.S. and M.S. degrees, all in electrical engineering, from the Universidad Simón Bolivar, Carcas, Venezuela, in 1987 and 1990, respectively. From 1992 to 1995, she worked with the Robust Speech Processing Laboratory and received the Ph.D. in biomedical engineering from Duke University, Durham, NC, in 1995.

She became an Assistant Professor of Electrical Engineering at the Universidad Simon Bolivar in 1995. Her current interests are digital signal processing of physiological signals in applications that address problems in electrocardiography and speech pathology.

James F. Kaiser (S'50–A'52–SM'70–F'73) was born in Piqua, OH, in 1929. He received the electrical engineering degree from the University of Cincinnati, Cincinnati, OH, in 1952 and the S.M. and Sc.D. degrees in 1954 and 1959, respectively, from Massachusetts Institute of Technology, Cambridge, all in electrical engineering.

He is currently Visiting Professor in the Department of Electrical and Computer Engineering at Duke University, Durham, NC, and Visiting Professor in the Department of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ. He was formerly a Distinguished Member of Technical Staff in the Speech and Image Processing Research Division of Bell Communication Research, Inc. since 1984. Prior to that he was a Member of Technical Staff at Bell Laboratories, Murray Hill, NJ for 25 years. He worked in the areas of speech processing, vocal tract modeling, system simulation, digital signal processing, computer graphics, and computer-aided design. He is the author of more than 70 research papers and the co-author and editor of eight books in the signal processing and automatic control areas.

Dr. Kaiser is a Fellow of the AAAS, a Registered Engineer in Massachusetts, and also a member of the ASA, EURASIP, and SIAM societies. He is a member of the honor societies Sigma Xi, Tau Beta Pi, and Eta Kappa Nu. He has served in a number of positions in both the Signal Processing Society and the Circuits and System Society.