Acoustic Analysis of Aperiodic Voice: Perturbation and Nonlinear Dynamic Properties in Esophageal Phonation

*Julia K. MacCallum, †Li Cai, †Liang Zhou, *Yu Zhang, and *Jack J. Jiang, *Madison, Wisconsin and †Shanghai, China*

Summary: Esophageal voice is a method of voice production after total laryngectomy. Previous research suggests that perturbation analysis may inaccurately measure aperiodic voices and that nonlinear dynamic methods may be more appropriate for analyzing signals of this type. Therefore, we hypothesized that nonlinear dynamic analysis would be more capable than perturbation parameters for reliable measurement of the aperiodic esophageal voice. The study design was acoustic comparison of esophageal and normal voice cohorts using nonlinear dynamic and perturbation measures. Twenty subjects in two age-matched groups participated in the study. Jitter, shimmer, signal-to-noise ratio (SNR), correlation dimension, and second-order entropy were measured from audio recordings of subjects' voices. Jitter and shimmer values were significantly higher and SNR values were significantly lower for esophageal voices than for normal voices. Error (err) count values, which indicate perturbation analysis reliability, were 0 in normal voices and significantly higher in esophageal voices. Error was attributable to signal aperiodicity and demonstrated that perturbation analysis yielded questionable results for esophageal voice. However, nonlinear dynamics measures analyzed both cohorts reliably and indicated that esophageal voice was significantly more chaotic than normal voice. The results demonstrated the capability of nonlinear dynamic methods to reliably quantify both aperiodic and periodic signals and differentiate normal from esophageal voices. It is suggested that nonlinear dynamic analysis be used preferentially for acoustic characterization of aperiodic voices, such as esophageal voice. Future research should focus on clarification of perturbation analysis reliability and further application of nonlinear dynamic measures to aperiodic voices.

Key Words: Aperiodic voice–Esophageal voice–Perturbation analysis–Nonlinear dynamic analysis–Percentage jitter–Percentage shimmer–Signal-to-noise ratio–Correlation dimension–Second-order entropy.

INTRODUCTION

Total laryngectomy is the removal of the entire larynx, including the thyroid and cricoid cartilages, the upper tracheal rings, and the hyoid bone.[1] This total excision of laryngeal tissue drastically affects both respiration and phonation in the patient and typically requires rehabilitative training. A number of postsurgical methods of voice restoration are available. Tracheoesophageal puncture (TEP) is the preferred method of voice restoration postlaryngectomy in the United States, the United Kingdom, and the continental Europe[2]; other speech methods, such as electrolarynx and esophageal voice, are applied in a limited manner. Although usage of TEP is dominant in certain North American and European countries, esophageal voice remains a significant laryngectomee voice production method in Asian countries, such as China. Jacobson et al noted the difficulties of TEP speech attainment in non-English/French speakers.[3] It is possible that language-related issues may result in the usage of alternative methods of voice restoration in China. Esophageal voice is generated using the esophagus as an air supply and the pharyngoesophageal segment, located on the superior aspect of the esophagus, as a vibration source.[4] Air is deposited in the esophagus, and its controlled release leads to segment vibration and sound production.[5] Speech is produced by moving the articulators to mouth words in the normal manner.[6]

The esophageal voice demonstrates perceptual qualities of harshness, hoarseness, gurgling, short phonation duration, and low pitch and volume,[5] all of which are indicators of aperiodicity in voice. Perturbation analysis of this perceptual abnormality has been accomplished, with results confirming the severe irregularity of esophageal voices. Signal-to-noise ratio (SNR), fundamental frequency, and intensity values were found to be lower and jitter and shimmer values were found to be higher in esophageal speech than in laryngeal speech.[4,5,7] However, recent research suggests that the validity of perturbation analysis is highly questionable when these measures are applied to aperiodic signals.[8,9]

In contrast, nonlinear dynamic methods of acoustic analysis have shown potential to reliably quantify both periodic and aperiodic signals. The concept of human voice production as a chaotic system has been established in recent years through computer modeling, excised larynx experiments, and human voice analysis.[10–14] A severely chaotic voice often exhibits an irregular and aperiodic waveform, poor perceptual qualities, and extreme perturbation values. Because all human vocal folds exhibit some inherent chaotic properties (eg, chaotic elements resulting from the nonlinear pressure-flow relation in the glottis, the nonlinear stress-strain interaction of vocal fold tissue, and the nonlinearities of vocal fold collision[11]), nonlinear dynamic methods are useful for quantifying the degree of aperiodicity and irregularity.

Previously, nonlinear dynamic analysis had not been applied to esophageal voices. Nor had the effectiveness of nonlinear

Accepted for publication October 2, 2007.

From the *Department of Surgery, Division of Otolaryngology Head and Neck Surgery, University of Wisconsin Medical School, Madison, Wisconsin; and the †Department of Otorhinolaryngology, Eye, Ear, Nose & Throat Hospital, Fudan University, Shanghai, China.

Address correspondence and reprint requests to Li Cai, Department of Otorhinolaryngology, Eye, Ear, Nose & Throat Hospital, Fudan University, 83 Fen Yang Road, Shanghai 200031, China. E-mail: dr.licai@yahoo.com

Journal of Voice, Vol. 23, No. 3, pp. 283-290
0892-1997/$36.00

doi:10.1016/j.jvoice.2007.10.004

dynamic analysis and perturbation analysis when quantifying the highly aperiodic esophageal voice been addressed. In this study, alaryngeal esophageal voice was compared to normal voice using perturbation and nonlinear dynamic analyses. It was hypothesized that nonlinear dynamic parameters would measure the highly aperiodic signals of esophageal voice more capably than perturbation measures, while still effectively distinguishing between esophageal and normal voice.

METHODS

Subjects

The Institutional Review Board at Shanghai EENT Hospital approved the protocol and consent procedure used in this study. Ten male speakers with esophageal voices participated in the study. All esophageal voice subjects were native speakers of Mandarin Chinese ranging in age from 50 to 65 years (mean = 55.4). Esophageal speakers underwent total laryngectomy for cancer of the larynx and received esophageal voice training to improve postsurgical vocal quality. In addition to the esophageal speakers, 10 age-matched male Mandarin Chinese speakers (age range = 42–60, mean age = 50.7) with normal voices were included for comparison. The normal voice subjects were healthy volunteers with no current or past evidence of voice disorders and normal larynges, as determined by clinical examination performed by an otolaryngologist.

Recording procedure

Audio recordings of the 10 laryngectomized patients' voices were collected at an average of 5.78 months after commencement of esophageal voice training (SD = 4.27). At each recording session, both normal and esophageal speakers were instructed to sustain phonations of /a/ at a comfortable pitch and volume for as long as possible. Recordings were made in a soundproof room with the recording microphone (AKG Acoustics, Vienna, Austria) positioned 10 cm from the subject's mouth. Audio files were recorded at a sampling rate of $f_s = 44.1$ kHz using the Multi-Dimensional Voice Program, Model 3650 (Kay Elemetrics Corp., Lincoln Park, NJ). A middle stationary segment $x(t_i),\ t_i = i\Delta t,\ \Delta t = 1/f_s,\ i = 1, 2\ldots,$ with a length of 1 second was selected for analysis from each subject's recording. Voice onset and offset were excluded to avoid effects of speech intonations or interactions between the larynx and vocal tract on analyses. Perturbation and nonlinear dynamic measures were then applied to these normal and esophageal voices.

Perturbation analysis

Perturbation analysis was conducted on the voice segments with *CSpeech* software, version 4.0 (Milenkovic and Read, Madison, WI). Two common measures of perturbation are jitter and shimmer. Jitter measures the cycle-to-cycle frequency variation of a voice signal, whereas shimmer measures the cycle-to-cycle amplitude variation. SNR, measured in decibels, reflects the dominance of the harmonic signal over aperiodic noise. Percentage jitter, percentage shimmer, and SNR values were obtained for each voice segment using *CSpeech*.

Reliability of perturbation measurements was determined by using error in perturbation measurement (err), a parameter calculated by the *CSpeech* program. Err counts "the number of times the analysis algorithm failed to compute a pitch period consistent with the peak of the autocorrelation function."[15] The autocorrelation function is used to calculate percentage jitter, percentage shimmer, and SNR values in *CSpeech*[16]; therefore, err describes the validity of all three measures. Algorithm failure can be attributed to inaccurate pitch estimation or a failed attempt to analyze a highly aperiodic waveform.[15] In comparing the err count of signals, Karnell et al found that signals with err > 0 produced significantly higher perturbation values and were more aperiodic than signals with err = 0.[9] An err count of 0 indicates that the analysis algorithm did not fail to compute pitch period intervals, meaning that error was sufficiently low and the signal was analyzed reliably via perturbation methods. Although minor algorithm failure corresponding to err = 0 is considered acceptable, err > 10 indicates loss of pitch tracking[15] and indicates the unreliability of perturbation results.

Nonlinear dynamic analysis

Detailed descriptions of nonlinear dynamic analysis methods, such as phase space reconstruction, correlation dimension, and second-order entropy to human voice production are widely found in the literature.[10–14,17–20] These types of measures have complemented traditional perturbation measurements because of an ability to describe chaotic, aperiodic voices.[10,11,13]

A reconstructed phase space can be created by plotting a voice signal against itself at some time delay. The reconstructed phase space qualitatively shows the dynamic behavior of a signal, as a periodic signal produces a closed trajectory whereas an aperiodic signal appears irregular.[21] Correlation dimension D_2 quantifies the irregularity of the reconstructed phase space: $D_2 = 0$ corresponds to a static state; $D_2 = 1$ corresponds to a periodic oscillation; $D_2 = 2$ describes a quasi-periodic tori (superposition of two or more oscillations with no rational fundamental frequencies or biphonation), and fractal D_2 describes an aperiodic or chaotic oscillation.[22] Kolmogorov entropy quantifies the rate of loss of information about the state of a dynamic system as it evolves.[23] Second-order entropy (K_2) indicates the rate of loss of information about the state of a dynamic system over time. For periodic behavior, this entropy is equal to zero. A chaotic system with a finite degree of freedom has a finite K_2 value, whereas the K_2 value of true random behavior approaches infinity.[23]

In this study, correlation dimension and second-order entropy calculations were performed using *Nonlinear Dynamic Analysis* software developed by the Laryngeal Physiology Laboratory at the University of Wisconsin. Calculations made by the software were based on the numerical algorithms described for studies analyzing excised larynx phonations[10] and pathological human voices.[13,19,20] Briefly, an m-dimensional delay-coordinate phase space $X_i = \{x(t_i), x(t_i - \tau), \ldots, x(t_i - (m-1)\tau)\}$ was reconstructed using the time delay technique,[21] where m is the embedding dimension and τ is the time delay. Dimension m was