An Objective Parameter to Classify Voice Signals Based on Variation in Energy Distribution

Boquan Liu, Evan Polce, and Jack Jiang, *Madison, Wisconsin*

Summary: Objectives. The purpose of this paper is to introduce an iterative nonlinear weighted method based on the variation in spectral energy distribution present in a voice signal to differentiate between four voice types: type 1 voice signals are nearly periodic, type 2 voice signals have strong modulations and subharmonics, type 3 signals are chaotic, and type 4 signals are dominated by stochastic noise.

Study Design. A total of 135 voice signal samples of the sustained vowel /a/ were obtained from the Disordered Voice Database and then individually categorized into the appropriate voice types based on the classification system described in Sprecher et al (2010). Voice samples were analyzed using the nonlinear methods of spectrum convergence ratio, rate of divergence, and nonlinear energy difference ratio (NEDR) to investigate classifier efficacy.

Methods. An iterative nonlinear weighted method based on the derivative of instantaneous frequency and Fourier transformations is applied to calculate spectral energy distributions. The distribution is then used to calculate the NEDR to classify voice signal types.

Results. Statistical analysis revealed that NEDR effectively differentiated between all four voice types ($P < 0.001$). Subsequent multiclass receiver operating characteristic analysis demonstrated that NEDR (area under the curve [95% CI] = 0.99 [0.96−1.0]) possessed the greatest classification accuracy relative to spectrum convergence ratio and rate of divergence.

Conclusion. NEDR was shown to be an effective metric for objective differentiation between all four voice signal types. NEDR calculations occurred approximately instantaneously, constituting a substantial improvement over the tedious computational time required for calculation of previous nonlinear parameters. This metric could assist clinicians in the diagnosis of voice disorders and monitor the efficacy of treatment through observation of voice acoustical improvement over time.

Key Words: Nonlinear weighted−Derivative of instantaneous frequency−Chaos−Voice signal classification−Nonlinear energy difference ratio.

INTRODUCTION

Voice disorders caused by physical injury, voice misuse, physiological disease, and surgery can impart significant functional and psychological limitations on the lives of patients.[1] Therefore, a more comprehensive understanding of the acoustic basis of voice disorders is imperative to properly diagnose and treat patients. To characterize voice disorders effectively, Titze[2] developed a classification scheme that assigned voice signals into three signal types based on visual interpretation of spectrograms. Type 1 voice signals are nearly periodic, meaning that the generated spectrograms display clearly defined harmonics and fundamental frequencies that appear nearly straight. Type 2 voice signals exhibit strong modulations and subharmonics. In type 2 spectrograms, strong modulations refer to the condition where the harmonics appear undulated rather than straight, whereas strong subharmonics refer to the presence of subharmonic frequencies that consist of interharmonic noise with intensities approaching the strength of the harmonics. Type 3 signals are characterized by chaotic dynamics with a finite dimension. This visual classification scheme was modified by Sprecher et al[3] to include a fourth voice type, which primarily exhibits stochastic noise behavior. Although spectrogram evaluation is capable of distinguishing between the four voice types, perceptual-based evaluation can be analytically inefficient and inconsistent at times due to its arbitrary nature as well as discrepancies in internal definitions, experience, and rating criteria between different subjective evaluators.[4−7] In acoustic analysis research, however, perceptual spectrogram evaluation is a useful gold standard comparison for developing and training objective acoustical classification methods. Compared to spectrogram evaluation, acoustic analysis is a more advantageous option clinically because it is objective, easier to comprehensively standardize, and computationally efficient.

Various linear and nonlinear acoustical analysis techniques have been previously utilized to classify voice signals into their respective category. Linear parameter-based perturbation analysis such as jitter and shimmer are only capable of classifying nearly periodic type 1 voice signals.[3] These measurements are deduced from calculations of the fundamental frequency and peak amplitude of each phonatory cycle; however, when phonation is characterized by substantial aperiodicity, jitter and shimmer are unable to produce accurate estimates. Therefore, jitter and shimmer are unreliable metrics for the analysis of type 2, type 3, and type 4 voice signals.[3]

On the contrary, nonlinear dynamic measurements have been shown to effectively quantify the differences between normal and irregular phonations. These methods include Kolmogorov

Accepted for publication February 14, 2018.
Conflict of interest: None.
From the Department of Surgery-Division of Otolaryngology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin.
Address correspondence and reprint requests to Jack Jiang, University of Wisconsin School of Medicine and Public Health, Department of Surgery-Division of Otolaryngology, 1300 University Avenue, 2745 Medical Sciences Center, Madison, WI 53706. E-mail: jiang@surgery.wisc.edu
Journal of Voice, Vol. 33, No. 5, pp. 591−602
0892-1997

https://doi.org/10.1016/j.jvoice.2018.02.011

entropy, correlation dimension (D2), and largest Lyapunov exponent.[8–11] Kolmogorov entropy describes the rate of information loss in a dynamic system.[10] D2 was first applied in the field of voice acoustics to analyze the dimensionality of cries produced by newborn infants,[12] and it represents the number of degrees of freedom required to describe the complexity of a system. Lyapunov exponents are used to characterize the average exponential rates of divergence or convergence of infinitesimally close orbitals in phase space of a dynamical system.[11] However, when the self-similarity property of a signal is destroyed by noise, for example, by aspiration due to turbulence in the vocal tract, these nonlinear methods fail to accurately identify voice type. Consequently, none of the aforementioned parameters are reliable objective measurements for the classification of all four types of voice signals.[1,13,14]

To differentiate more effectively between all four types of voice signals, Lin et al[15] and Calawerts et al[16] introduced spectrum convergence ratio (SCR) and rate of divergence (ROD), respectively. SCR uses short-time Fourier transform to quantify the convergence of 250 generated segments for each voice signal. Briefly, a time series obtained from a voice signal is segmented by a windowing function that moves along the time axis. The window size determines the number of sampled points and is generally set to 0.012 seconds to generate 250 constituent segments for each signal. Then Fourier transformations (FTs) are performed for each of the 250 segments. Discrete short-time Fourier transforms are commonly used to analyze discrete segments from time series to determine changes in frequency across the entire signal. The SCR can then be defined to quantify the similarity, or convergence, of constituent frequencies across the segments composing a signal. The SCR relies on the assumption that periodic voice signals are composed of extremely similar frequency segments, whereas aperiodic voice signals contain considerably dissimilar segments; however, the SCR can be insensitive to small frequency changes in voice signals due to inadequate frequency resolution, resulting in computational errors. The ROD utilizes a modified Wolf algorithm to calculate Lyapunov exponents, but when large amounts of noise are present in the signal, ROD analysis breaks down.[15,16]

Additionally, investigating the relationship between nonlinear dynamic analysis and subjectively perceived voice features is crucial to further elucidating the applicability of these parameters. Roughness, breathiness, strain, and overall dysphonia severity are common auditory-perceptual features of voice that previous studies have attempted to correlate with the results of objective acoustic analysis.[17] Perceptions of phonation breathiness and roughness are often due to turbulent airflow stemming from incomplete glottal closures and abnormal vocal fold muscle tension and mucosal membrane elasticity, respectively.[17,18] The resulting phonatory signals are characterized by substantial aperiodicity, but, in theory, should be quantifiable by measures of periodicity, such as nonlinear dynamic analyses. Previously, research involving cepstral-based acoustic measures has indicated that cepstral peak prominence (CPP) is correlated with perceived overall dysphonia severity, breathiness, and strain[19]; however, a weaker correlation exists between CPP and phonatory roughness.[18] Poor correlations with roughness suggest that the perception of roughness may not be intrinsically related to changes in periodicity, but rather an entirely different aspect of voice signals. On the contrary, fewer studies have focused on elucidating the relationship between nonlinear dynamic analysis and the subjective perception of voice features. Moderate strength correlations have been demonstrated between overall dysphonia severity and the nonlinear parameters of D2[13] and Lyapunov exponents.[20] However, future investigations are warranted to illuminate the correlational relationship between nonlinear dynamic metrics and the perceived voice features of roughness, breathiness, and strain.

In this paper, a nonlinear energy difference ratio (NEDR) is presented to identify and distinguish between all four voice signal types. In signal processing, FTs are employed to decompose time signals into their constituent frequencies. The energy of a signal can then be calculated through summation of the spectral energies of the frequency components according to Parseval theorem.[21] Based on a nonlinear weighted method, which has been widely used in sensor networks and image processing, the proposed method applies the derivative of instantaneous frequency (IF) to establish an iterative algorithm for spectral energy variation calculation.[22–25] The employed nonlinear weighted function utilizes a moving window to weigh local data points based on their relative position to the nth data point, with data points in close proximity to the nth data point weighted more heavily. Much like weights can be incorporated into a sample mean to form a weighted mean, the data points surrounding the nth data point of interest are more heavily weighed to improve the accuracy of spectral energy distribution calculation, as well as augment the overall energy resolution.

NEDR is designed to characterize the time-varying features of nonstationary signals. Accordingly, a signal that displays strong periodicity (type 1) exhibits stable spectral energy distribution. Whereas, if a signal is breathy or aperiodic (types 3 and 4), the spectral energy distribution can vary considerably. We defined NEDR to quantify this spectral energy variation and classify voice signals into one of the four corresponding voice type categories. The efficacy of the proposed method is demonstrated by comparison with SCR and ROD.

METHODS

Iterative algorithm for voice types classification

An iteration step is adopted to obtain the spectral energy distribution of a voice signal x, $x = [x_1, x_2 \ldots x_N]$, and $N = f_s \times t$ is the number of sample data points used in the calculation, f_s and t are sampling frequency and time length of x, respectively. Figure 1 provides a visual flow chart of the sequential steps taken for computation of the NEDR.

At the i+1 step, a square matrix X^{i+1} is as follows:

$$X^{i+1} = \begin{bmatrix} Y_{1,1}^{i+1} & Y_{1,2}^{i+1} & \cdots & Y_{1,N}^{i+1} \\ Y_{2,1}^{i+1} & Y_{2,2}^{i+1} & \cdots & Y_{2,N}^{i+1} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{N,1}^{i+1} & Y_{N,2}^{i+1} & \cdots & Y_{N,N}^{i+1} \end{bmatrix}, \quad (1)$$

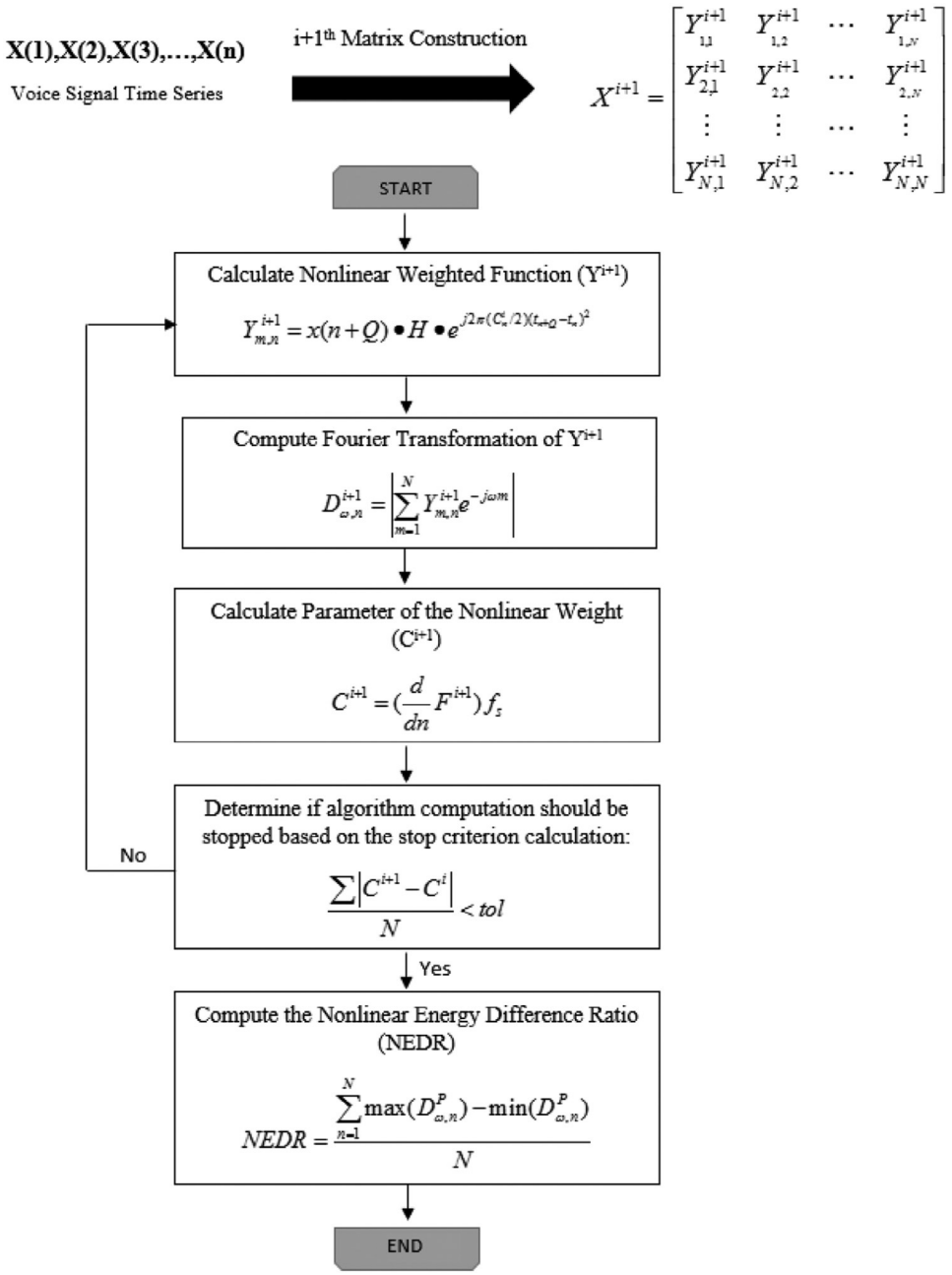

FIGURE 1. Flow chart demonstrating the computational steps involved in calculating the nonlinear energy difference ratio (NEDR). First, an i +1 square matrix is constructed by calculating a nonlinear weighted function (Y^{i+1}) from voice signal time series data x, where $x = [x_1, x_2 \ldots x_N]$. Second, Fourier transformations (FT) were performed to decompose the voice signal into the frequency domain and obtain spectral energy and frequency distributions. Next, the nonlinear weight parameter (C^{i+1}) is calculated to determine if the algorithm should stop and proceed to computation of the NEDR or perform subsequent iterations of the previous steps to improve accuracy. The stop criterion compares the difference between the value of the nonlinear weight parameter in the current iteration (C^{i+1}) with the nonlinear weight parameter from the previous iteration (C^{i}). If the stop criterion is satisfied, then the iterative algorithm is stopped and the NEDR is calculated. The entire computational process of NEDR calculation for one voice sample takes approximately 20 seconds to complete.

where

$$Y_{m,n}^{i+1} = \begin{cases} x(n+Q)\cdot H\cdot e^{j2\pi(C_n^i/2)(t_{n+Q}-t_n)^2} & m = rem(N+Q,N)+1 \\ 0 & else \end{cases} \quad m,n = 1,2,\ldots,N, \tag{2}$$

$$H = real(M) - j \cdot imag(M), \tag{3}$$

$$M = W(P/2 + Q + 1), \tag{4}$$

where $Q = [-min([round(N/2)-1, P/2, n-1]) \ldots min([round(N/2)-1, P/2, N-n])]$, the $t_n = n/f_s$ is the nth time point, the $W = [W_1, W_2, \ldots W_P]$ stands for a window function, the P is the window length used, $rem(N+Q,N)$ is the remainder after division of $N+Q$ by N, $real(\cdot)$ and $imag(\cdot)$ define the real and imaginary parts of a complex number, respectively, $H \cdot e^{j2\pi(C_n^i/2)(t_{n+Q}-t_n)^2}$ represents the nonlinear weight, and Y^{i+1} is the nonlinear weighted function, as discussed earlier, of the voice signal x in the i+1 iteration. The nonlinear weight is utilized for sampling of data points in the time series $x = [x_1, x_2 \ldots x_N]$. In a time series obtained from observational voice signal data, the data points in close proximity to the data point of interest are closely related to and provide crucial information about the nth data point. Thus, the nonlinear weight calculation ensures that data adjacent to the nth data point are more heavily weighted.

The FT is used to obtain the spectral energy distribution and frequency.

$$D_{\omega,n}^{i+1} = \left| \sum_{m=1}^{N} Y_{m,n}^{i+1} e^{-j\omega m} \right|. \tag{5}$$

The $|\bullet|$ stands for the absolute value. The frequency value corresponding to each sample point is recorded in F^{i+1}. Following the FT, a voice signal in the time domain is decomposed into its constituent frequency components in the frequency domain. This decomposition generates frequency data over the course of the entire signal and, thus, to distinguish frequency data from different parts of the signal, an arbitrary digital time is employed. The rate at which the signal frequency changes with respect to digital time is then designated as the derivative of IF ($\frac{d}{dn}F^{i+1}$). The $C^{i+1} = [C_1^{i+1}, C_2^{i+1}, \ldots, C_{N-1}^{i+1}, C_{N-1}^{i+1}]$ is a parameter of the nonlinear weight in the i+1 iteration, C_n^{i+1} is the nth element in the C^{i+1}.

$$C^{i+1} = \left(\frac{d}{dn} F^{i+1} \right) f_s. \tag{6}$$

The criterion used to determine whether the algorithm should be stopped is as follows:

$$\frac{\sum |C^{i+1} - C^i|}{N} < tol. \tag{7}$$

tol is a value controlling the end of the iterations. In this paper, the *tol* value is 10. Specifically, Equation (7) is used to measure the change in C^i between successive iterations, that is, differences in the weights of successive computations of the algorithm. To optimize computational running time, the algorithm calculation is stopped if the change in C^i is small. We have determined through pilot studies with NEDR that a *tol* value of 10 operates as a suitable balance point for optimization of computational time without compromising the accuracy of the spectral energy distribution calculation. Therefore, a *tol* value of 10 was determined to be sufficient for the purposes of this study.

The NEDR is defined to quantify variation in the spectral energy distribution across time. The NEDR can be calculated when the stop criterion is reached:

$$NEDR = \frac{\sum_{n=1}^{N} \max(D_{\omega,n}^P) - \min(D_{\omega,n}^P)}{N}, \tag{8}$$

the $D_{\omega,n}^p$ is the value of $D_{\omega,n}^i$ when the stop criterion is reached. In this paper, the $C^0 = [0, 0, \ldots 0]$. Based on the mathematical equations discussed earlier, a custom *MATLAB* R2017a (MathWorks Natick, MA) program was created for NEDR calculation.

Voice selection

One hundred thirty-five samples of the vowel /a/ were randomly selected and analyzed from the Disordered Voice Database Model 4337 KayPENTAX (Lincoln Park, NJ). The methodology employed for collecting the patient voice data in the KayPENTAX database was standardized for all voice recordings. The voice samples were recorded by a condenser microphone located 15 cm away from the subjects' mouths in a sound-attenuated booth. The subjects phonated a sustained /a/ vowel into the microphone for one second and were recorded using a sampling frequency of 44.1 kHz. Once the voice recordings were randomly selected from the database, the samples were cut to a length of 0.75 second to match the signal length input requirements of the custom *MATLAB* algorithms employed in this study. The 0.75-second segments were selected based on visual inspection and subjective determination of the most stable segment of the voice waveform, excluding vocal onset and offset. All samples used in analysis were recorded from different subjects. Summary characteristics for the samples used in analysis are displayed in Table 1.

TABLE 1.
Subject Characteristics

Voice Type	Number of Samples	Age in Years	Gender
1	34	36.7 (19–81)	9 men 23 women
2	35	38.8 (17–73)	6 men 11 women
3	42	43.5 (18–80)	14 men 16 women
4	24	62.5 (40–93)	5 men 9 women

Notes: age and gender information is displayed for the 93 voice samples for which patient information was provided. Forty-two of the voice samples included in this study did not have diagnostic or patient information provided by KayPENTAX. Age is displayed as mean age (age range).

Spectrogram analysis

A Hamming window shape was used to generate narrowband spectrograms for each voice sample. A window length of 50 milliseconds, time step of 0.002 second, frequency step of 5 Hz, and dynamic range of 40 dB were utilized for the creation of each spectrogram. The spectrograms generated for the voice samples were then individually categorized into their respective voice types based on the classification system presented in Sprecher et al.[3] Three previously employed researchers trained in acoustic phonetics and spectral analysis performed the spectrogram analysis task by assigning a subjectively determined voice type to each of the 135 voice samples. The researchers were blinded and the spectrograms were randomized for each voice sample. If all three of the researchers' spectrogram ratings were not equivalent, the spectrograms with conflicting ratings were rerandomized and the researchers were tasked with performing spectrogram analysis on this subset of spectrograms a second time. After a second round of individual spectrogram classifications, all spectrograms that still were given incongruent ratings were then subjectively analyzed by the three researchers in conjunction. Samples that were atypical representations of the voice types or classifications that were unable to be agreed upon by all three of the researchers after group discussion were excluded from analysis. This process was used to ensure that incorrect subjective classification of the voice samples was minimized.

Due to inter-rater and intrarater reliability statistics not being performed during initial spectrogram analysis, we enlisted and trained two laboratory researchers to individually classify a randomly selected subset of 30 of the 135 total voice spectrograms that were used in this study. A spectrogram sample size of 30 was determined large enough to produce an accurate approximation of the inter- and intrarater reliability of the spectrogram analysis paradigm used in this study. Training for the spectrogram evaluation task was identical to the training completed by the three previously employed researchers and consisted of one session of preliminary instruction outlining the basic definitions and characteristics of the four different voice type spectrograms as described in Sprecher et al.[3] Briefly, type 1 spectrograms exhibit strong periodicity, clearly defined and straight harmonics, and minimal interharmonic noise. Type 2 spectrograms display harmonic modulations, bifurcations, and subharmonic frequencies. Type 3 spectrograms are characterized by diffuse energy smearing that obscures the harmonic frequencies above 1500 Hz. Type 4 spectrograms often do not have a clearly defined fundamental frequency and contain a more pervasive smearing of energy across frequencies relative to type 3 spectrograms.

The spectrogram classification results from the two trained researchers were compared to assess classification reliability among different judges. Additionally, 10 of the 30 spectrograms were repeated during subjective evaluation to determine intrarater consistency. Results from the inter-rater reliability assessment indicated that the perceptual designations made by the two trained judges agreed for 90% (27/30) of the spectrogram classifications. Similarly, the results of the intrarater reliability assessment indicated a 95% (19/20) combined accuracy for the two judges. These results indicate a high level of classification agreement within and between judges following spectrogram analysis training. The classification consistency exhibited in spectrogram analysis was considered adequate for the purposes of this study.

Statistical analysis

To compare the results of NEDR with SCR and ROD, calculations of standard deviation within each voice type, one-way analysis of variance (ANOVA) between different voice types, and a multiclass receiver operating characteristic (ROC) curve for each method were performed. Subsequently, area under the curve (AUC) and 95% confidence interval (CI) values were calculated to compare the classification performances of each method. If the ANOVA was significant ($P < 0.001$), pairwise *t* tests were performed to discern which groups possessed significantly different mean values. A d-prime statistic was utilized to analyze the effect sizes between the means of the voice types. Effect sizes measure and quantify the difference or separation between two group means, and a larger effect size indicates a more robust difference between the mean values of two groups.[26] Generally, a d-prime value greater than 0.8 is considered to indicate a large size effect. Scatter plots and box plots were generated to depict the data.

The amplitudes of the data distribution curves for NEDR, SCR, and ROD displayed in Figures 1–3 were normalized and smoothed to allow for clear comparisons between the locations of the distribution curves that correspond to each of the four voice types.

RESULTS

Voice type classification

As shown in Figure 2, the NEDR values increased as voice type increased. The NEDR values within the same voice type fluctuated within a narrow range around the mean, demonstrating the stability of the proposed method for an individual voice type (Figure 2). Compared with SCR and ROD, the location of the normalized distribution curves of NEDR for each voice type exhibited reduced overlap.

Figures 3 and 4 demonstrate that the results of SCR and ROD analysis within the same voice type exhibited increased variability and deviation from the mean value. The data distribution curves of SCR and ROD values for each voice type can be visualized in Figures 3 and 4, respectively. The SCR and ROD values exhibited greater scattering away from the mean for each voice type relative to NEDR, which is evident from the greater degree of overlapping between the voice type distribution curves. SCR values decreased with increasing voice type, whereas ROD values increased.

One-way ANOVA

One-way ANOVA results showed that significant differences existed between the means of the voice types for all three nonlinear methods. Results from the pairwise *t* tests and d-prime analyses for NEDR, SCR, and ROD are displayed in Tables 2–4,

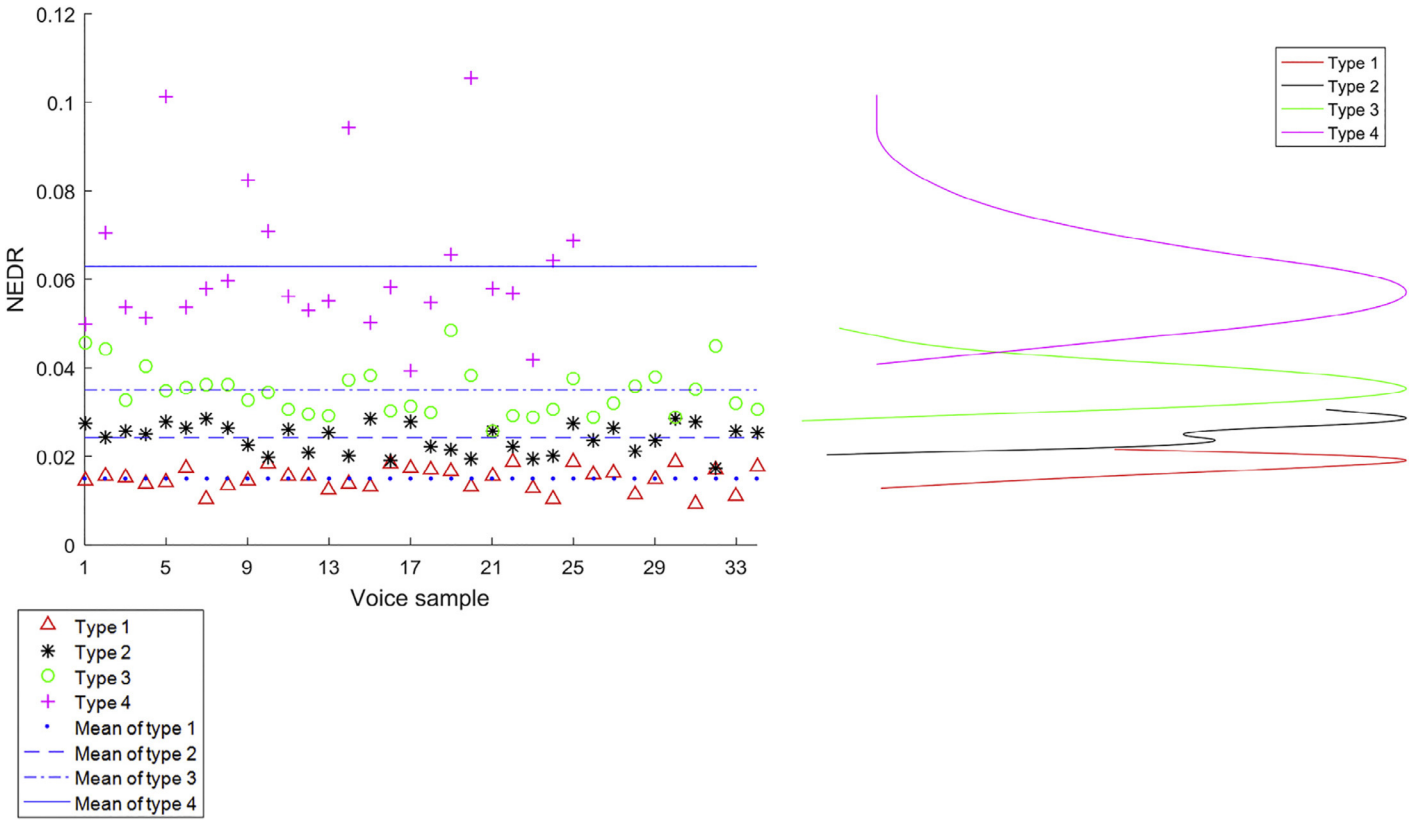

FIGURE 2. The nonlinear energy difference ratio (NEDR) values and normalized data distribution for different voice types from 135 recorded voice samples. The data corresponding to each of the four voice types are represented by a specific shape and color. Mean NEDR values are depicted as *lines* on the scatter plot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

respectively. NEDR was the only metric that yielded significant mean differences ($P < 0.001$) and substantial effect sizes ($D' > 0.8$) between the means of all four voice type groups, which can be seen in Table 2 and visualized in Figure 5A.

SCR analysis yielded significant differences for type 1 versus type 3, type 1 versus type 4, type 2 versus type 4, and type 3 versus type 4, but no significant differences for the remaining pairs. Additionally, d-prime statistical analysis indicated substantial size effects between the mean values of type 1 versus type 4, type 2 versus type 4, and type 3 versus type 4, but exhibited moderate to small effect sizes for the remaining comparisons (Figure 5B and Table 3).

The primary limitation of the ROD method was observed overlapping between the ROD values of type 1, type 2, and type 3 voices (Figure 5C). ROD analysis yielded significant differences for all pairwise comparisons except type 2 versus type 3; however, d-prime statistical analysis revealed that substantial effect sizes were only found for type 1 versus type 3, type 1 versus type 4, type 2 versus type 4, and type 3 versus type 4 comparisons (Table 4).

TABLE 2.
NEDR Comparisons Between Groups

Comparison	P	D′
Type 1 vs. type 2	<0.001	1.56
Type 1 vs. type 3	<0.001	2.50
Type 1 vs. type 4	<0.001	2.46
Type 2 vs. type 3	<0.001	1.25
Type 2 vs. type 4	<0.001	1.92
Type 3 vs. type 4	<0.001	1.25

Note: pairwise *t* tests and d-prime statistics were used to analyze if NEDR was effective at distinguishing between the four voice types.
Abbreviation: NEDR, nonlinear energy difference ratio.

TABLE 3.
SCR Comparisons Between Groups

Comparison	P	D′
Type 1 vs. type 2	0.110	0.19
Type 1 vs. type 3	<0.001	0.54
Type 1 vs. type 4	<0.001	2.14
Type 2 vs. type 3	0.006	0.38
Type 2 vs. type 4	<0.001	1.87
Type 3 vs. type 4	<0.001	0.95

Note: pairwise *t* tests and d-prime statistics were used to analyze if SCR was effective at distinguishing between the four voice types.
Abbreviation: SCR, spectrum convergence ratio.

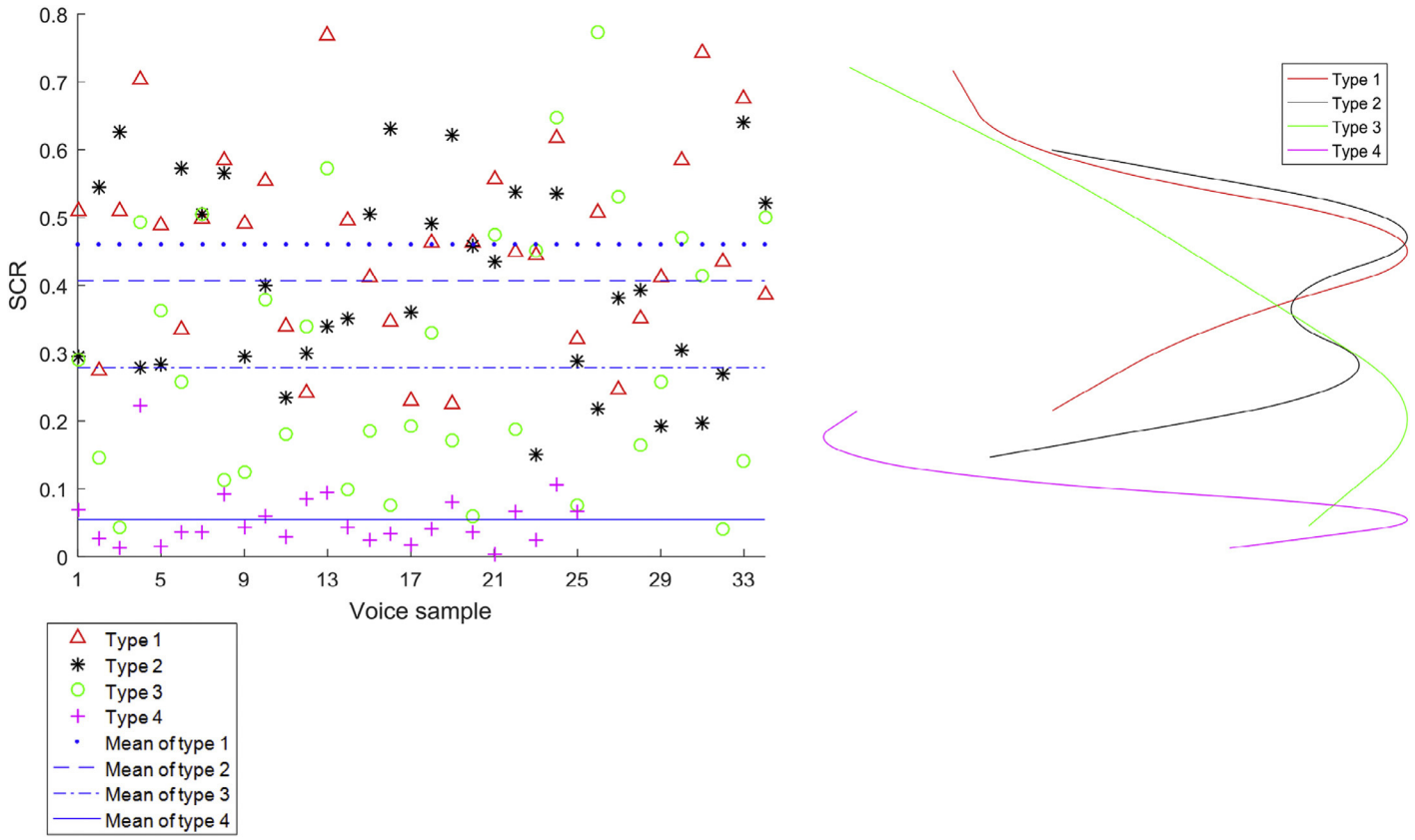

FIGURE 3. The spectrum convergence ratio (SCR) values and normalized data distribution for different voice types from 135 recorded voice samples. The data corresponding to each of the four voice types are represented by a specific shape and color. Mean SCR values are depicted as *lines* on the scatter plot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ROC

One hundred thirty-five samples were utilized to construct multiclass ROC curves for each method (Figure 6). The four classes used to determine classification efficacy were the type 1 and type 2 voice samples, type 2 and type 3 voice samples, type 3 and type 4 voice samples, and type 4 voice samples compared to the other three voice type samples. The calculated AUC and 95% CI values for NEDR, SCR, and ROD are displayed in Table 5. For all classification conditions analyzed, NEDR exhibited the largest AUC relative to SCR and ROD. To provide a single metric for comparison between different classification methods, the four AUC values for each method were averaged to generate a mean AUC value. The mean AUCs and 95% CIs were 0.99 (0.96−1.0), 0.80 (0.70−0.88), and 0.83 (0.73−0.90) for NEDR, SCR and ROD, respectively.

TABLE 4.
ROD Comparisons Between Groups

Comparison	P	D′
Type 1 vs. type 2	<0.001	0.49
Type 1 vs. type 3	<0.001	0.81
Type 1 vs. type 4	<0.001	2.33
Type 2 vs. type 3	0.030	0.31
Type 2 vs. type 4	<0.001	1.53
Type 3 vs. type 4	<0.001	1.09

Note: pairwise *t* tests and d-prime statistics were used to analyze if ROD was effective at distinguishing between the four voice types.
Abbreviation: ROD, rate of divergence.

DISCUSSION

In this paper, the proposed method of NEDR was compared with the current nonlinear metrics of SCR and ROD. SCR analysis was effective in quantifying type 4, but failed to accurately distinguish between type 1 and type 2 voice. The ROD method had limited capabilities in classifying type 1, type 2, and type 3 voice signals. Only NEDR demonstrated comprehensive classification of all four voice types effectively.

Figure 2 demonstrates that the obtained NEDR data for each voice type were highly localized to discrete ranges of NEDR values that, consequently, yielded significantly different means and large effect sizes between the voice types. Contrastingly, increased variability and fluctuations in the SCR and ROD values obtained within and between each voice type (Figure 3 and Figure 4, respectively) indicated that these two metrics might be unreliable for accurately differentiating between all four voice types. Construction of multiclass ROC curves and calculated AUC values allowed for a direct comparison of overall classifier performance between the methods (Figure 6). NEDR exhibited the greatest mean AUC value

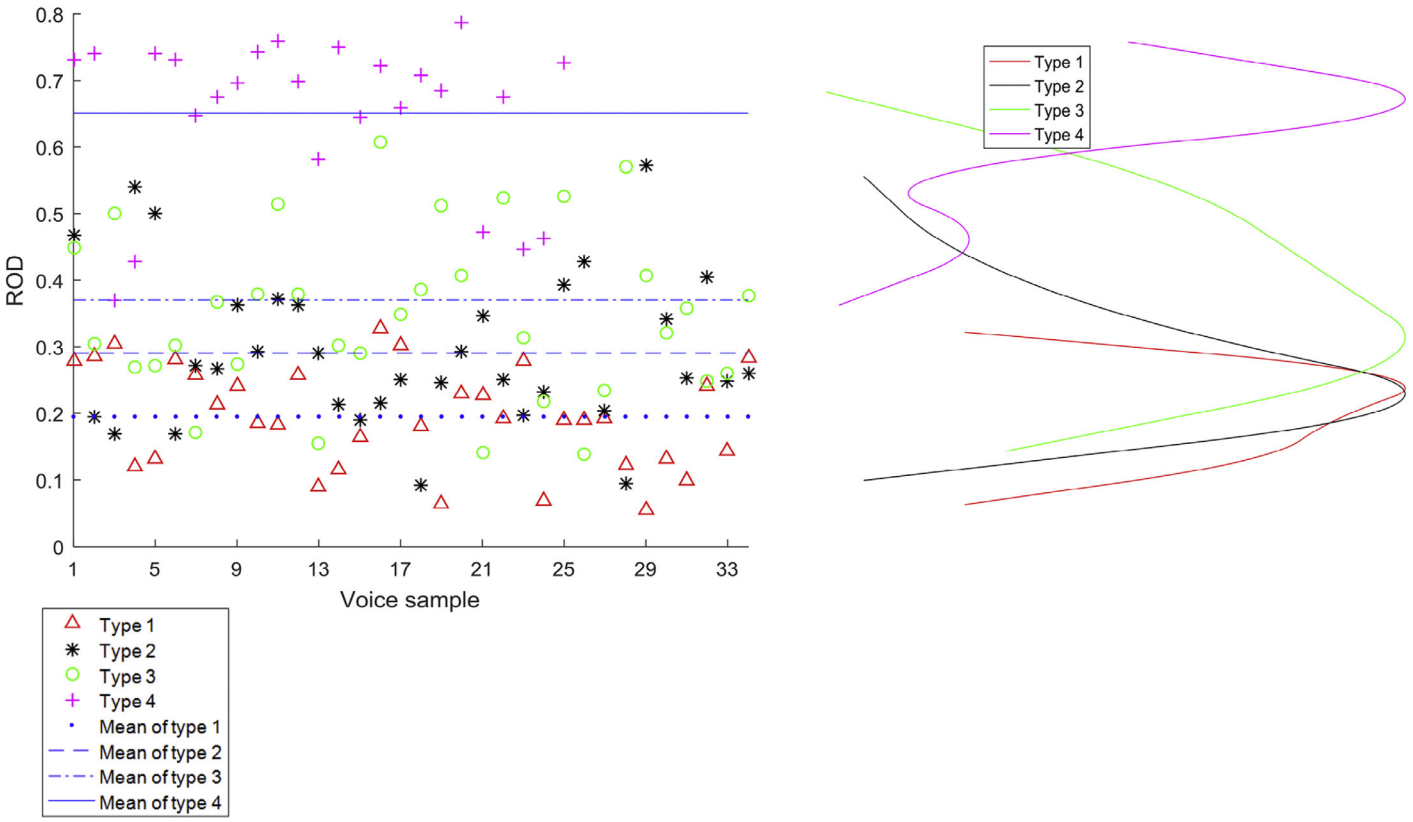

FIGURE 4. The rate of divergence (ROD) values and normalized data distribution for different voice types from 135 recorded voice samples. The data corresponding to each of the four voice types are represented by a specific shape and color. Mean ROD values are depicted as *lines* on the scatter plot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(0.99 [0.96–1.0]) and, consequently, demonstrated the most robust classification efficacy of the methods analyzed.

Determination of voice type from acoustic output provides a noninvasive, objective tool that can enable clinicians to obtain quantitative data pertaining to functional voice production. Rabiner and Schafer[27] explained a model of voice production where vocal fold vibrations generate a quasiperiodic excitation signal in the vocal tract, which comprises most of the energy in a voice signal. Recently, this traditional voice production model has been modified to incorporate the four-voice type scheme. Voice type signals 1, 2, and 3 comprise the low-dimensional vibratory system and are produced by periodic and chaotic vibrations of the vocal folds, whereas type 4 voice signals originate from nonlinear stress-strain interaction of vocal fold collisions and tissues and the resulting turbulent airflow through the vocal tract.[11,28,29] Although all voice signals contain some degree of infinitely dimensional noise due to turbulence, the stochastic noise component of type 4 signals is strong and pervasive; as a result, stochastic noise dominates over low-dimensional components. Investigation of the functional differences between type 3 and 4 voices is clinically relevant because it can generate insight into the fundamental airflow dynamics and biomechanical interactions present in disordered phonation. Thus, differentiation between type 3 and 4 voice types can

TABLE 5.
AUC Values and 95% CIs for NEDR, SCR, and ROD Obtained From Multiclass ROC Analysis

Comparison	NEDR	SCR	ROD
Type 1 and type 2	0.99 (0.96–1.0)	0.59 (0.45–0.72)	0.74 (0.63–0.85)
Type 2 and type 3	0.99 (0.94–1.0)	0.72 (0.59–0.83)	0.68 (0.54–0.79)
Type 3 and type 4	0.99 (0.95–1.0)	0.92 (0.83–0.97)	0.93 (0.83–0.97)
Type 4 and all others	0.99 (0.99–1.0)	0.97 (0.93–0.99)	0.96 (0.92–0.99)
Average	0.99 (0.96–1.0)	0.80 (0.70–0.88)	0.83 (0.73–0.90)

Note: results are presented as AUC (95% CI).
Abbreviations: AUC, area under the curve; CI, confidence interval; NEDR, nonlinear energy difference ratio; ROC, receiver operating characteristic; ROD, rate of divergence; SCR, spectrum convergence ratio.

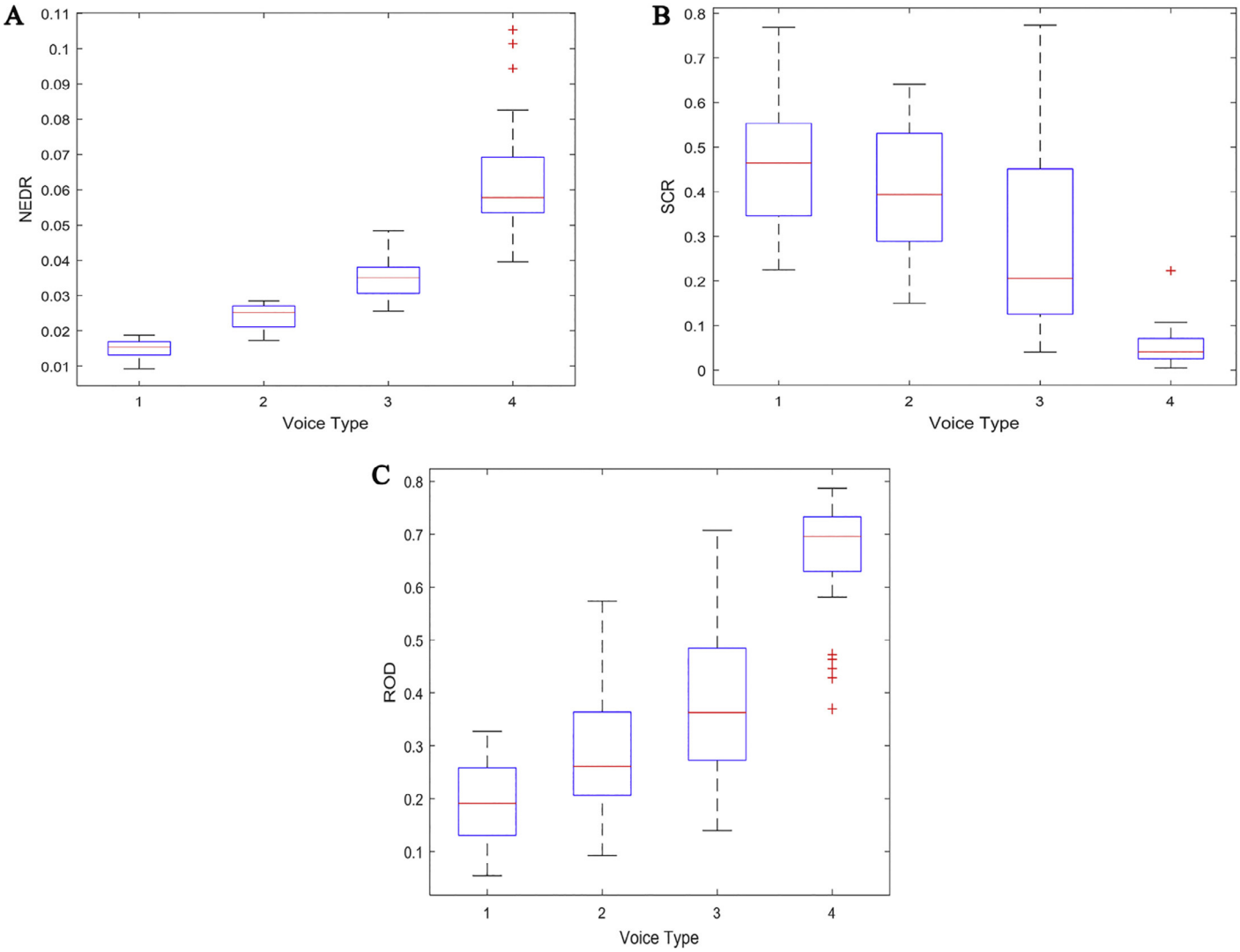

FIGURE 5. Box-and-whisker plots of (**A**) nonlinear energy difference ratio (NEDR) of all four voice types, (**B**) spectrum convergence ratio (SCR) of all four voice types, and (**C**) rate of divergence (ROD) of all four voice types. Voice type is designated along the x-axis. Boxes represent interquartile ranges. Median values are indicated by the *solid red lines*. The *red pluses* represent outliers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

provide valuable acoustic information leading to further investigations into the functional and morphological abnormalities inherent in disordered voice production.

Nonlinear phenomena in voice production are not only observed in disordered voice, but also in normal, nonpathological aspects of speech. Although current research is focused on classifying voices into discrete signal type categories, in reality, human phonation is not static, but rather exhibits nonstationary behavior and is comprised of varying degrees of periodic and aperiodic components. Thus, voice production is frequently modeled and thought of as a series of nonlinear coupled oscillators that generate complex, nonlinear phenomena such as subharmonic frequencies, period doubling bifurcations, toroidal representations in phase space, and chaotic vibrations.[30–32] Even strongly periodic type 1 voice signals display evidence of low-dimensional chaotic characteristics.[30] The extent to which nonlinear phenomena and chaotic vibratory patterns predominate over periodic components governs the overall dynamics and characteristics of the voice signal. Accordingly, previous studies have demonstrated that nonlinear phenomena in normal voice production can be employed to augment artistic individuality by musical performers and reinforce social groups.[31,33] Specifically, the vocal fry register is perceptually characterized as an aperiodic phonation with a lower pitch and creaky phonation aspect that has gained popularity with female adolescent populations. Acoustically, vocal fry exhibits considerable aperiodicity in both signal duration and amplitude. However, vocal fry is not considered a voice disorder, but rather a typical component of normal voice production. Similarly, acoustic analysis of vocalized singing has indicated that musical performers can induce various nonlinear vocal phenomena, such as sonorities, subharmonics, biphonations, and chaos. By definition, the four voice type paradigm simply represents descriptive classes of common voice signal characteristics. Vocal fry and musical vocalizations contain nonlinear signal components that pertain to characteristics of

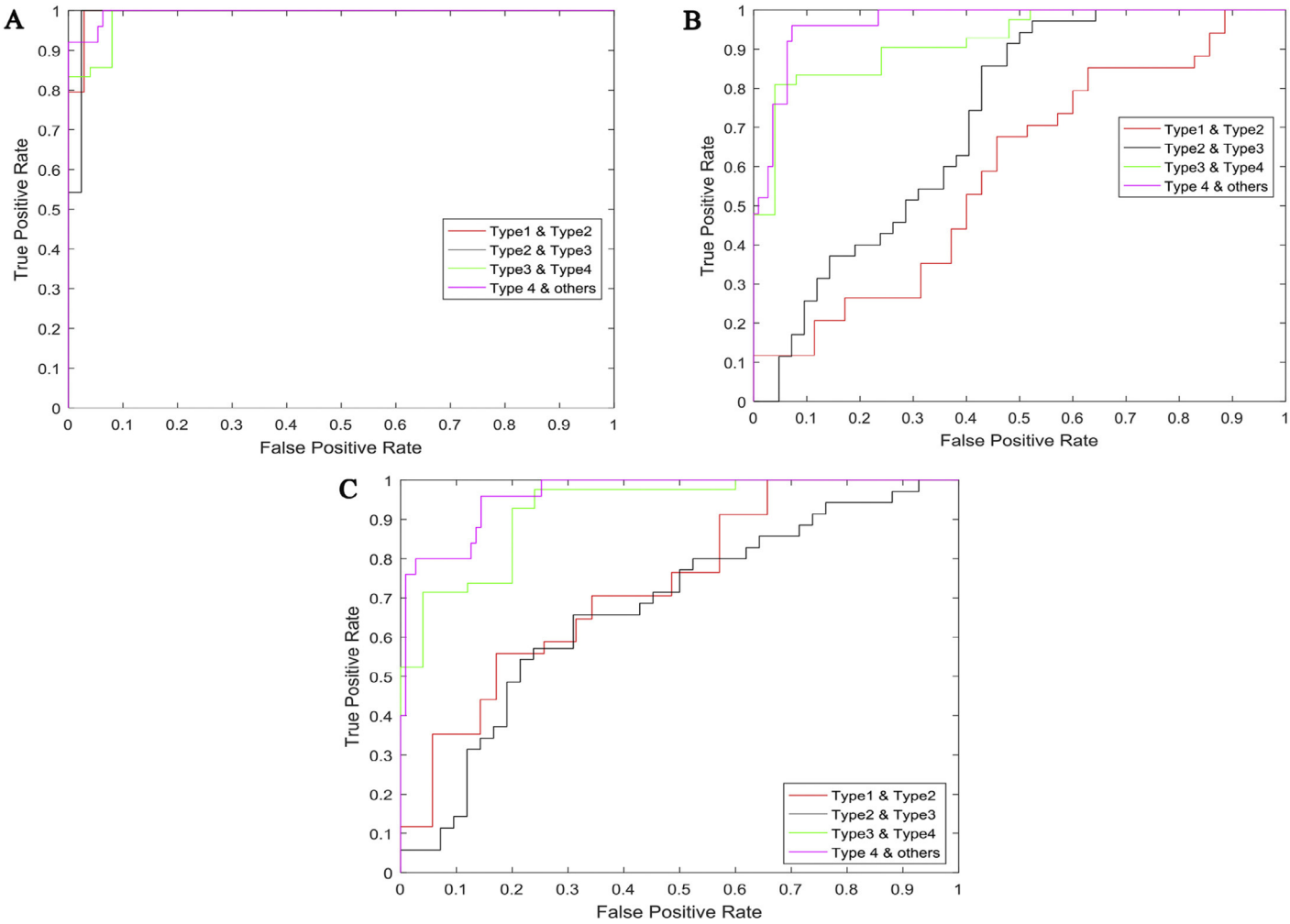

FIGURE 6. ROC plots for classifier performance comparison. (**A**) Nonlinear energy difference ratio (NEDR), (**B**) spectrum convergence ratio (SCR), (**C**) rate of divergence (ROD).

type 3 and 4 voice; however, a type 3 or 4 designation in this case does not necessarily suggest that the signal represents disordered phonation. Rather, it implies that the signal has certain spectral characteristics indicating that chaotic or stochastic components govern the signal dynamics. Thus, certain voice production styles can be characterized by substantial chaos and aperiodicity; however, spectral disorder does not necessarily always indicate underlying pathology.

Although the classification efficacy of NEDR in the current study is promising, limitations of this nonlinear method will need to be addressed in future research. As previously discussed, the spectral energy distribution of predominantly periodic signals exhibits stable energy distribution, whereas the spectral energy distribution of breathy, aperiodic signals is characterized by considerable variation. Other recording factors that influence spectral energy distribution include environmental noise, which can degrade recorded voice signal quality,[34] as well as intra- and intersubject variation in loudness and pitch. Generally, noise in recorded voice samples can be distinguished as either stationary noise, or constant environmental noise, or nonstationary noise, such as background sounds in the clinical or research setting that vary unpredictably during the recording process. Filtering stationary noise from recorded speech is not difficult to accomplish; however, nonstationary noise within the frequency range of human phonation can be much more problematic to filter. As such, nonstationary background noise can potentially augment the variation in spectral energy distribution, which would artificially inflate the calculated NEDR value. In the current study, the obtained voice samples were recorded in a sound-proof booth, which might not be representative of recording conditions in clinical offices. Additionally, the various frequencies and amplitudes composing a voice signal contribute to the signal energy; thus, intra- and intersubject variability in loudness and pitch during voice recording is expected to impact the spectral energy distribution as well. Currently, it is common in acoustical analysis research for subjects to phonate at a self-determined comfortable loudness and pitch,[35] which makes investigating the effect that loudness and pitch have on voice classification rather ambiguous. Future studies will need to examine the effect that recording voices in a variety of environments, such as in quiet and noisy office settings, as well as intra- and intersubject variability in loudness and pitch has on NEDR output and classification.

Another potential limitation of the current study is that NEDR was only applied for analysis of the sustained /a/ vowel. Acoustic analysis currently utilizes recording of multiple sustained vowels (/a/, /e/, /i/, and /u/).[36,37] Articulation of different vowels necessitates several physical alterations including differing tongue and laryngeal cartilage positions, glottal areas, and vocal fold tensions. Previous studies investigating the effect of vowel selection on the computational output of traditional acoustic analysis parameters have observed varying results. The findings of some studies have indicated that lower values of jitter and shimmer were observed in higher vowels,[38] whereas other studies observed the contrary.[39,40] Regarding the effect of vowel selection on the computation of NEDR, the maximum and minimum spectral energy values that are specific to low and high vowel waveforms differ to some degree and, therefore, future studies should focus on investigating and establishing voice type boundary values for different vowels. However, the classification performance of NEDR, relative to SCR and ROD, is not expected to deteriorate when analyzing different sustained vowels.

Unlike sustained vowels, connected speech comprised characteristics such as speaking rate, dialect, intonation, and articulation, and more closely resembles the dynamic aspects of normal speech.[41] However, extending the application of acoustic analysis to connected speech has proven difficult because traditional perturbation methods are adversely affected by inconsistencies in pitch and loudness, noise stemming from the consonants, and shorter vowel lengths. Zhang and Jiang[42] investigated the acoustic characteristics of sustained and running vowels from normal and disordered subjects and suggested that traditional perturbation parameters, such as jitter and shimmer, might not be appropriate for the analysis of connected speech. In contrast, cepstral measures, including CPP, are not predicated on frequency tracking and have been shown to be capable for analysis of connected speech.[43] Similarly, studies have also demonstrated that nonlinear dynamic methodology can be applied for analysis of running vowels and fricative consonants.[42,44] An assessment of the capability of NEDR for analysis of the phonatory characteristics of connected speech is necessary to determine the sensitivity of NEDR to these vocal conditions and tasks.

Experimental results demonstrated that the proposed method has a direct relationship with voice type. The NEDR values were lowest in type 1 voice signals and increased with voice signal type, which is a consequence of increasing interference in the voice signal. The computational time for the custom *MATLAB* program to calculate NEDR was approximately 20 seconds for each voice sample. For the purposes of this research, NEDR demonstrated proficiency in producing virtually instantaneous data. Although the NEDR calculation is efficient, other aspects of acoustic analysis research involved in this study, such as visual classification of spectrograms, are still rather tedious. Future studies will focus on creating NEDR voice type value boundaries to eliminate the inefficiencies of spectrogram classification and for optimization of real-time classification of observational voice data. In conclusion, the proposed NEDR parameter serves as an effective technique to quantify the broad spectral energy differences introduced by interference in voice signals.

CONCLUSION

In summary, the NEDR is a nonlinear weighted method based on the derivative of IF that generates an iterative algorithm for calculation of spectral energy variation in a signal. The normalized data distributions of NEDR values from within each voice type exhibited reduced variability and fluctuation when compared with SCR and ROD. Subsequently, one-way ANOVA, pairwise t test, and d-prime statistical analyses revealed that NEDR produced the most significant difference between the means of each voice type. Accordingly, ROC and AUC analysis demonstrated the proficiency of the proposed method in voice type classification of all four types of voice signals. Future studies could utilize NEDR to analyze the spectral energy differences present in common voice disorders and then generate boundaries of NEDR values where specific voice disorders typically are observed.

Currently, the utility of acoustic analysis in clinical practice is limited by deficiencies in the computational speed and accuracy of objective parameters; however, the results of this study suggest that NEDR might offer a solution to the shortcomings of previous methods. The computation of NEDR approximately occurred in real-time, constituting a substantial improvement over the laborious and time-intensive requirements of previous nonlinear methods. The NEDR could assist clinicians in obtaining additional acoustical information, which may lead to superior treatment for individuals suffering from voice disorders. Furthermore, by measuring NEDR changes over the course of treatment interventions, quantitative evaluation of treatment and disease progression is possible.

Acknowledgments

This study was supported and funded by National Institutes of Health, NIH Grant No. R01-DC006019 from the National Institute on Deafness and Other Communication Disorders.

REFERENCES

1. Little MA, McSharry PE, Roberts SJ, et al. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomed Eng Online*. 2007;6:23.
2. Titze IR. Workshop on acoustic voice analysis: summary statement. *National Center for Voice and Speech*. 199526–27. Denver, CO; Available at: http://www.ncvs.org/freebooks/summary-statement.pdf.
3. Sprecher A, Olszewski A, Zhang Y, et al. Updating signal typing in voice: addition of type 4 signals. *J Acoust Soc Am*. 2010;127:3710–3716.
4. Gerratt BR, Kreiman J. Measuring vocal quality with speech synthesis. *J Acoust Soc Am*. 2001;110:2560–2566.
5. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am*. 1998;104:1598–1608.
6. Millet B, Dejonckere PH. What determines the differences in perceptual rating of dysphonia between experienced rater? *Folia Phoniatr Logop*. 1998;50:305–310.
7. Gupta R, Chaspari T, Kim J, et al. *Pathological speech processing: state-of-the-art, current challenges, and future directions*. IEEE Conference Proceedings. 20162016:6470–6474.

8. Awan SN, Novaleski CK, Rousseau B. Nonlinear analyses of elicited modal, raised, and pressed rabbit phonation. *J Voice*. 2014;28:538–547.
9. Choi SH, Zhang Y, Jiang JJ, et al. Nonlinear dynamic-based analysis of severe dysphonia in patients with vocal fold scar and sulcus vocalis. *J Voice*. 2012. https://doi.org/10.1016/j.jvoice.2011.09.006.
10. Jiang JJ, Zhang Y, Ford CN. Nonlinear dynamics of phonations in excised larynx experiments. *J Acoust Soc Am*. 2003;114(4 pt 1):2198–2205.
11. Jiang JJ, Zhang Y, McGilligan C. Chaos in voice, from modeling to measurement. *J Voice*. 2005;20:2–17.
12. Mende W, Herzel H, Wermke K. Bifurcations and chaos in newborn infant cries. *Phys Lett*. 1990;145:418–424.
13. Awan SN, Roy N, Jiang JJ. Nonlinear dynamic analysis of disordered voice: the relationship between the correlation dimension (D2) and pre-/post-treatment change in perceived dysphonia severity. *J Voice*. 2010;24:285–293.
14. Ma EP, Yiu EM. Suitability of acoustic perturbation measures in analysing periodic and nearly periodic voice signals. *Folia Phoniatr Logop*. 2005;57:38–47.
15. Lin L, Calawerts WM, Dodd K, et al. An objective parameter for quantifying the turbulent noise portion of voice signals. *J Voice*. 2016;30:664–669.
16. Calawerts WM, Lin L, Sprott JC, et al. Using rate of divergence as an objective measure to differentiate between voice signal types based on the amount of disorder in the signal. *J Voice*. 2017;31:16–23.
17. Herzel H, Reuter R. Quantifying correlations in pitch- and amplitude contours of sustained phonation. *Acta Acustica United Acustica*. 2000;86:129–135.
18. Heman-Ackah YD, Michael DD, Goding GS. The relationship between cepstral peak prominence and selected parameters of dysphonia. *J Voice*. 2002;16:20–27.
19. Sauder C, Bretl M, Eadie T. Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and Analysis of Dysphonia in Speech and Voice (ADSV). *J Voice*. 2017;31:557–566.
20. Yu P, Ouaknine M, Revis J, et al. Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *J Voice*. 2001;15:529–542.
21. Oppenheim AV, Schafer RW. *Discrete-Time Signal Processing*. 2nd ed Upper Saddle River, NJ: Prentice Hall; 1999.
22. Chen J, Li J, Yang S, et al. Weighted optimization-based distributed Kalman filter for nonlinear target tracking in collaborative sensor networks. *IEEE Trans Cybern*. 2016;99:1–14.
23. Liu B, Zeng Y. Uncertainty-aware frequency estimation algorithm for passive wireless resonant SAW sensor measurement. *Sens Actuators A Phys*. 2016;237:136–146.
24. Rafajlowicz E, Pawlak M, Steland A. Nonlinear image processing and filtering: a unified approach based on vertically weighted regression. *Int J Appl Math Comput Sci*. 2008;18:49–61.
25. Shmaliy YS. Suboptimal FIR filtering of nonlinear models in additive white Gaussian noise. *IEEE Trans Signal Process*. 2012;60:5519–5527.
26. Kelley K, Preacher KJ. On effect size. *Psychol Methods*. 2012;17:137–152.
27. Rabiner LR, Schafer RW. *Digital Processing of Speech Signals*. Upper Saddle River, NJ: Prentice Hall; 1978.
28. Zhang Y, Jiang JJ, Biazzo L, et al. Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis. *J Voice*. 2005;19:519–528.
29. Zhang Y, McGilligan C, Zhou L, et al. Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps. *J Acoust Soc Am*. 2004;115:2270–2277.
30. Tao C, Jiang JJ. Chaotic component obscured by strong periodicity in voice production system. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2008;77 061922.
31. Neubauer J, Edgerton M, Herzel H. Nonlinear phenomena in contemporary vocal music. *J Voice*. 2004;18:1–12.
32. Kumar A, Mullick SK. Nonlinear dynamical analysis of speech. *J Acoust Soc Am*. 1996;100:615–629.
33. Abdelli-Beruh NB, Drugman T, Red Owl RH. Occurrence frequencies of acoustic patterns of vocal fry in American English speakers. *J Voice*. 2016;30:759 e11-759.e20.
34. Aronsson C, Bohman M, Ternstrom S, et al. Loud voice during environmental noise exposure in patients with vocal nodules. *Logoped Phoniatr Vocol*. 2007;32:60–70.
35. Awan SN, Roy N. Outcomes measurement in voice disorders: application of an acoustic index of dysphonia severity. *J Speech Hear Res*. 2009;52:482–499.
36. Higgins MB, Netsell R, Schulte L. Vowel-related differences in laryngeal articulatory and phonatory function. *J Speech Hear Res*. 1998;41:712–724.
37. Moon KR, Chung SM, Park HS, et al. Materials of acoustic analysis: sustained vowel versus sentence. *J Voice*. 2012;26:563–565.
38. MacCallum JK, Zhang Y, Jiang JJ. Vowel selection and its effects on perturbation and nonlinear dynamic measures. *Folia Phoniatr Logop*. 2011;63:88–97.
39. Kiliç MA, Öğüt F, Dursun G, et al. The effects of vowels on voice perturbation measures. *J Voice*. 2004;18:318–324.
40. Orlikoff RF. Vocal stability and vocal tract configuration: an acoustic and electroglottographic investigation. *J Voice*. 1995;9:173–181.
41. Parsa V, Jamieson DG. Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *J Speech Hear Res*. 2001;44:327–339.
42. Zhang Y, Jiang JJ. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *J Voice*. 2008;22:1–9.
43. Awan SN, Roy N, Jett ME, et al. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the CAPE-V. *Clin Linguist Phon*. 2010;24:742–758.
44. Narayanan SS, Alwan AA. A nonlinear dynamical systems analysis of fricative consonants. *J Acoust Soc Am*. 1995;97:2511–2524.