An Objective Parameter for Quantifying the Turbulent Noise Portion of Voice Signals

Liyu Lin, William Calawerts, Keith Dodd, and Jack J. Jiang, *Madison, Wisconsin*

Summary: Objectives. Currently, there are no objective measures capable of distinguishing between all four voice signal types proposed by Titze in 1995 and updated by Sprecher in 2010. We propose an objective metric that distinguishes between voice signal types based on the aperiodicity present in a signal.

Study design. One hundred fifty voice signal samples were randomly selected from the Disordered Voice Database and subjectively sorted into the appropriate voice signal category on the basis of the classification scheme presented in Sprecher 2010.

Methods. Short-time Fourier transform was applied to each voice sample to produce a spectrum for each signal. The spectrum of each signal was divided into 250 time segments. Next, these segments were compared to each other and used to calculate an outcome named spectrum convergence ratio (SCR). Finally, the mean SCR was calculated for each of the four voice signal types.

Results. SCR was capable of significantly differentiating between each of the four voice signal types ($P < 0.001$). Additionally, this new parameter proved equally as effective at distinguishing between voice signal types as currently available parameters.

Conclusion. SCR was capable of objectively distinguishing between all four voice signal types. This metric could be used by clinicians to quickly and efficiently diagnose voice disorders and monitor improvements in voice acoustical signals during treatment methods.

Key Words: Turbulence–Signal spectrum analysis–Short time Fourier transform–Voice signal classification–Spectrum convergence ratio.

INTRODUCTION

In 1995, Titze proposed classifying voice signals into three signal types—type 1 voice signals are nearly periodic, type 2 voice signals have strong modulations and subharmonics, and type 3 signals are not periodic.[1] Type 1 and type 2 signals can be analyzed by perturbation parameters (jitter, shimmer). Nonlinear parameters, such as correlation dimension and second-order entropy, have been proven successful at differentiating between type 2 and type 3 voice signals.[2] In 2010, a new voice type to Titze's voice classification, type 4 voice, was introduced.[2] The difference between type 3 voice and type 4 voice signals in this scheme is that type 3 voice is chaotic with finite dimension, whereas type 4 voice is defined by severe breathiness and primarily stochastic noise characteristics. That is, the correlation dimension for type 3 voice signals converges to a specific value with increasing embedding dimension, whereas that of type 4 does not. Additionally, the spectrums for type 3 voice signals are characterized by energy centralization in lower frequencies, whereas type 4 voice signals exhibit a searing of energy across a broad range of frequencies.

Current linear parameters such as jitter and shimmer can classify type 1 and type 2 voice signals. Jitter represents the cycle-to-cycle variation in signal frequency and shimmer measures the cycle-to-cycle variation in signal amplitude.[2] Because these measurements are determined by estimating the fundamental frequency and peak amplitude of each phonatory cycle, they are unable to produce stable estimates for irregular phonation. Thus, they are neither valid nor reliable for analyzing type 3 and type 4 voice signals.

To combat this issue, Titze et al suggested that nonlinear parameters could quantify the difference between more complex voice signals. These parameters are Lyapunov exponents, correlation dimension (D2), and Kolmogorov entropy.[2,3] Lyapunov exponents, which are the average exponential rates of divergence or convergence of nearby orbits in phase space, are effective descriptors of chaos.[4] Thus, a higher Lyapunov exponent indicates that a system is more chaotic. Correlation dimension analysis calculates the number of degrees of freedom necessary to describe a system. A system with a higher degree of complexity requires more degrees of freedom to characterize its dynamic state.[4] Finally, Kolmogorov entropy is a description of the rate of information loss in a dynamic system.[5] A larger Kolmogorov entropy value indicates a more complex system.

Calculations of correlation dimension and Lyapunov exponents from excised larynx experiments demonstrate that low-dimensional chaotic behavior exists in phonation.[5] Furthermore, correlation dimension and second-order Kolmogorov entropy (K2) have been proven be useful in the analysis of sustained and running vowels.[4] However, when the signal is contaminated by a large amount of noise, for example, aspiration caused by turbulence in the vocal tract, nonlinear parameters break down. The turbulent energy in the vocal tract causes the signal to lose its self-similarity property, making these nonlinear calculations invalid.[6] Thus, nonlinear metrics such as D2, Kolmogorov entropy, and Lyapunov exponent cannot be calculated for this type of voice signal. Currently, only subjective measures are capable of distinguishing between type 3

Accepted for publication August 25, 2015.
Conflict of interest: None.

From the Division of Otolaryngology–Head and Neck Surgery, Department of Surgery, University of Wisconsin-Madison School of Medicine and Public Health, Madison, Wisconsin.

Address correspondence and reprint requests to Jack Jiang, 1300 University Avenue, 2745 Medical Sciences Center, Madison, WI 53706. E-mail: jjjiang@wisc.edu

Journal of Voice, Vol. ■, No. ■, pp. 1-6
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2015.08.017

and type 4 voice signals, making it difficult for researchers to establish criterion to classify voice signals in this scheme.

We reasoned that through using short-time Fourier transform (STFT) analysis, we could develop a continuous metric capable of distinguishing between all four types of voice signals. STFT is a powerful analysis tool for audio signal processing because it tracks how frequency components in a signal change with time.[7–9] Thus, by adjusting it to the proper time and frequency resolution, the transform is proved to be sensitive in detecting small changes in the periodicity of a signal. We examined each signal's spectrum because if the voice signal is affected by turbulent noise, the chaotic energy would deteriorate that spectrum's convergence. This is because a spectrum of a periodic system would consist of segments that closely resembled each other. Thus, as the complexity and aperiodicity of a system increases, the segments would resemble each other less. We developed a metric called spectrum convergence ratio (SCR) to quantify the degree that each segment resembled each other, or converged.

In this study, we hypothesized that SCR would be highest in type 1 voice signals and decreased as voice type increased. Additionally, we compared this metric to currently existing evaluation tools and hypothesized that SCR would be as effective at distinguishing between each voice signal type as these currently existing methods.

METHODS

Short-time Fourier transform

Because of the fact that some deterministic characteristics might be obscured by turbulent energy, we observed the signal's spectrum and time-frequency relationship to classify the signals into voice types. Traditionally, this is done by subjective classification.

Fourier analysis is a well-known tool in signal processing and aims to analyze the manifestation of time domain signals in the frequency domain and vice versa. The STFT is an extension of Fourier analysis. It defines a class of time-frequency distributions which specify complex amplitude versus time and frequency data for any signal.[9] Instead of analyzing the frequency components of the entire signal, the discrete Fourier transform is performed on segments of the signal, enabling the user to analyze changes (amplitude and phase) in frequency over time. STFT is commonly used to analyze voice signal's spectrums in the pattern recognition field. When applying STFT, the time sequence is divided into segments using a windowing function, and the Fourier transform of each segment is found. The discrete STFT is defined by

$$S_x(\omega,k)=\sum_{-\infty}^{+\infty}x(n)m(n-k)e^{-j\omega n}, \tag{1}$$

where $\mathrm{x(n)}$ is the time series and $\mathrm{m(n-k)}$ is the window function. At moment n, the window function reduces $\mathrm{x(n)}$ to zero outside a specified interval. As tag n moves along the time axis, the observing window is slid along the time axis, capturing local time segments. The result of this transform is a set of coefficients denoted by $\mathrm{S_x(\omega,k)}$.

Window size, which decides the number of sampling points in a local time segment, is an important factor in STFT. Different segment lengths produce different frequency and time resolutions. If the local time segment length is too small, frequency resolution will be poor, but if the length is too large, it will be difficult to analyze the details of changes in frequency.[10] In this study, a window size of 0.012 seconds was chosen, producing 250 segments for each sample.

Spectrum convergence ratio

Two hundred fifty spectrums were produced after applying STFT for each signal. Each segment was compared with the other segments by plotting their amplitudes as shown in Figure 1A. Under the assumption that the voice signal is a sustained vowel with a constant fundamental frequency, a signal that displays strong periodicity (type 1) would have segments that closely resemble each other. If a signal is breathy, or aperiodic (types 3 and 4), the segments should vary considerably from each other. We defined a variable called the dynamic range of segments' spectrogram (DRSS) to quantify the variation in frequency. By observing Figure 1A,C, we are able to distinguish type 1 and type 4 signals by measuring the area under the curve.

In a discrete model, the area can be calculated by

$$\mathrm{DRSS}=\sum[\mathrm{C_{max}(n)-C_{min}(n)}]\ , \tag{2}$$

where $\mathrm{C_{max}(n)}$ is the maximum-energy curve expression, whereas $\mathrm{C_{min}(n)}$ is the minimum-energy curve expression. They provide the maximum and minimum coefficients value in same time tag of all signal segments.

To find SCR, we first generated $\mathrm{S_x(\omega,n)}$ of a voice signal and recorded it into a matrix. In this matrix, each row is a spectrum of a segment, whereas each column containing the Fourier coefficients with same time tag in every segment. Next, we normalized each row by the maximum element in it and then plotted them to create a convergence graph. The difference between the maximal value and minimal value at every moment is the DRSS. We defined maximum energy (MAE) as

$$\mathrm{MAE}=\sum\mathrm{C_{max}(n)}. \tag{3}$$

Finally, the convergence ratio, which we named SCR, is found using the formula

$$\mathrm{SCR}=-\ln\left(\frac{\mathrm{DRSS}}{\mathrm{MAE}}\right). \tag{4}$$

Similar to jitter and shimmer, SCR is a parameter extracted from linear analysis results and is capable of analyzing signals with high-dimensional chaos turbulence. SCR comes from signal spectrogram analysis, but the methods to calculate DRSS and MAE of the signals' spectrogram were applied discrete integral and exponential calculation, making them nonlinear methods.

Correlation dimension (D2) analysis

Correlation dimension (D2) analysis is used to compare the metric proposed in this article to a metric that has already been