Published in final edited form as:

J Voice. 2016 November ; 30(6): 664–669. doi:10.1016/j.jvoice.2015.08.017.

An objective parameter for quantifying the turbulent noise portion of voice signals

Liyu Lin, PhD, William Calawerts, B.S., Keith Dodd, B.S., and Jack J. Jiang, M.D., PhD

University of Wisconsin-Madison School of Medicine and Public Health, Department of Surgery, Division of Otolaryngology- Head and Neck Surgery, Madison, WI, 53792

Abstract

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—150 voice signal samples were randomly selected from the Disordered Voice Database and subjectively sorted into the appropriate voice signal category based on 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. Lastly, the mean Spectrum Convergence Ratio was calculated for each of the four voice signal types.

Results—Spectrum Convergence Ratio 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—Spectrum Convergence Ratio was capable of objectively distinguishing between all 4 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.

Keywords

Turbulence; signal spectrum analysis; short time Fourier transform; voice signal classification; Spectrum Convergence Ratio

Corresponding author: Jack Jiang, M.D., Ph.D., 1300 University Avenue, 2745 Medical Sciences Center, Madison, WI 53706, Phone: (608) 265-9854, Fax: (608) 265-2139.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest: none

INTRODUCTION

In 1995 Titze proposed classifying voice signals into 3 signal types – type 1 voice signals are nearly periodic, type 2 voice signals have strong modulations and subharmonics, and type 3 signals are not periodic¹. 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 signals². In 2010, a new voice type to Titze's voice classification, type 4 voice, was introduced². The difference between type 3 voice and type 4 voice signals in this scheme is that type 3 voice is chaotic with finite dimension while 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, while that of type 4 does not. Additionally, the spectrums for type 3 voice signals are characterized by energy centralization in lower frequencies, while 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 amplitude². Since 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 non-linear 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 chaos⁴. 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 state⁴. Lastly, Kolmogorov entropy is a description of the rate of information loss in a dynamic system⁵. 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 phonation⁵. Furthermore, correlation dimension and second-order Kolmogorov entropy (K2) have been proven be useful in the analysis of sustained and running vowels⁴. However, when the signal is contaminated by a large amount of noise, for example aspiration caused by turbulence in the vocal tract, non-linear parameters break down. The turbulent energy in the vocal tract causes the signal to lose its self-similarity property, making these nonlinear calculations invalid⁶. 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 types 3 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 analysis we could develop a continuous metric capable of distinguishing between all four types of voice signals. Short-time Fourier Transform (STFT) is a powerful analysis tool for audio signal processing because it tracks how frequency components in a signal change with time⁷⁻⁹. 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 decrease 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

STFT (Short Time Fourier Transform)

Due to the fact that some deterministic characteristics might be obscured by turbulent energy, we observed the signal's spectrum and time-frequency relationship in order 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 Short-time Fourier Transform (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 signal⁹. 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:

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

Where $x(n)$ is the time series and $m(n-k)$ is the window function. At moment n , the window function reduces $x(n)$ 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 $S_x(\omega, k)$.

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 frequency¹⁰. In this study, a window size of 0.012 seconds was chosen, producing 250 segments for each sample.

SCR (Spectrum Convergence Ratio)

250 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 1(a). 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 1(a) and 1(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:

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

Where $C_{max}(n)$ is the maximum energy curve expression while $C_{min}(n)$ 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 $S_x(\omega, n)$ of a voice signal and recorded it into a matrix. In this matrix, each row is a spectrum of a segment while 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 MAE (Maximum Energy) as

$$MAE = \sum C_{max}(n) \quad (3)$$

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

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

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 paper to a metric that has already been proven effective. D2 quantifies the complexity of a system by measuring the number of degrees of freedom in a signal⁵. A time-delay embedding technique as described by Packard et al 1980 was used to calculate this value¹¹. The m-dimensional phase space of the acoustic signal was reconstructed by plotting m-dimensional vectors,

$$X_i = \{x(n_i), x(n_i - \tau), \dots, x(n_i - (m-1)\tau)\} \quad (5)$$

where $x(n)$ is the voice sample, τ is the time delay, and m is the minimum embedding dimension. To determine the appropriate time delay, the mutual information method was used¹². The box-counting algorithm was applied to estimate the correlation integral, Cr ¹³. Since the radius of the m-dimensional hypercube, r and Cr , is related by the following power law, $(r) \propto r^{D2} e^{-m\tau K2}$, D2 was found by calculating the slope of the most linear part of the $\log(Cr) \log(r)$ plot¹⁴⁻¹⁶.

A MATLAB nonlinear signal processing program named Open TSTool calculated the embedding dimension and D2. The most linear part of $\log(Cr)$ - $\log(r)$ curve was selected and the slope of this part was taken as the D2. Since D2 has been proven incapable of distinguishing between type 3 and type 4 voice signals, only type 1, 2 and 3 signals were analyzed in this study².

Statistical Analysis—150 samples were randomly selected and analyzed from the Disordered Voice Database Model 4337 KayPEN-TAX, Lincoln Park, NJ. The samples had a frequency rate of 25 kHz and the beginning and end portions were discarded to clip each sample to 0.75 seconds. Sample characteristics can be seen in Table I.

Then, a spectrogram was plotted for each sample. Using this spectrogram, three researchers independently sorted the samples into their respective voice types based on the classification scheme in Sprecher 2010². Samples which were not agreed upon by all three researchers, or samples that were not typical representations of a voice signal type were discarded. The process was repeated during a second round of classification, leaving a final sample set of 22 type 1 samples, 49 type 2 samples, 51 type 3 samples and 26 type 4 samples. This process was followed to minimize the risk of incorrect subject classification by the research team. For each voice type, a user-defined Short-Time-Fourier-Transform (STFT) function was used to produce the spectrum, and then the Spectrogram Convergence Ratio was found.

To compare the effectiveness of SCR to currently available metrics, a One Way ANOVA was run for D2, jitter, and shimmer. If the ANOVA was significant, pairwise t-tests were run to determine which groups had differences in the means using a Bonferroni correction, ($P < 0.002$).

The experiment was run with a window size of 0.012 seconds and a sample frequency of 25 kHz. We performed a One Way Analysis of Variance (ANOVA), ($P < 0.05$) on each voice type. If the ANOVA was significant, we ran pairwise t-tests to determine which groups had significant differences in the means using a Bonferroni correction, ($P < 0.0083$). Box plots and one-dimensional scatter plots were constructed to visualize the distribution of the data.

RESULTS

One way ANOVA

The results showed significant differences in the means of SCR for all voice types. This can be seen in Table II and visualized in Figure 2.

Jitter and shimmer analysis were proven effective in distinguishing between type 1 and type 2 voice signals. A box plot of this data can be seen in Fig. 3a,b. D2 analysis showed no significant differences in the mean for type 1 vs. type 2 voice signals, as is shown in Fig. 3c and Table III, but showed significant differences in the mean between type 2 and 3 voice types. D2 analysis was unable to quantify type 4 voice signals.

Figure 4 shows the histogram stochastic result of the SCR experiments. As expected for a periodic voice (type 1), the convergence ratio is high. For type 2 voices, the convergence ratio is moderately high because of frequency modulations. Thus, the DRSS is higher than type 1 voices, and the spectrum shows some variation in frequency over time, making the convergence graph show moderately high convergence. For type 3 voices signals, the convergence ratio is low because they do not have a periodic structure. The energy is smeared in low frequencies, and the DRSS is much larger than type 1 or type 2 voices. Lastly, the convergence was very low in type 4 voice signals and the spectrum showed much variation in frequency over time.

Performance comparison

Voice signal samples were taken as points in Cartesian coordinates, and the SCR was set as the x-axis and the y-axis was set to shimmer and D2 respectively. Then the results were plotted in Fig. 5. The performance of clustering in horizontal axis is better than in vertical axis for both of the subplots. In this way, SCR can be proven at least as effective as traditional measures of shimmer and D2.

DISCUSSION

In this paper, we compared the performances of SCR with traditional parameters that classify voice signal types. D2 analysis shows that there is a difference in the nonlinear dynamics of normal voices (types 1 and type 2) compared to type 3 but failed to characterize type 4 voices as supported by Sprecher 2010². Also, D2 failed to differentiate between type 1 and type 2 voice signals. The inability of D2 to characterize type 4 voices is due to the fact that type 4 voices contain a large breathy component, which destroys the self-similarity property of a chaotic system¹⁰. Type 4 signals differ from other voice types because they have a high percentage of turbulent noise. This reduces the SCR by enlarging MAE. Furthermore, SCR is able to classify all 4 types of signals effectively, implying that SCR has

the capability to detect the severity of frequency modulations. Experimental results show that SCR has an inverse relationship to voice type. SCR was highest in type 1 voice signals and decreased as voice signal type increased, which confirmed our hypothesis. Therefore, SCR offers a continuous variable that quantifies the periodicity of a system and allows for analysis of complex systems.

This proposed method attempts to quantify the smearing of energy across the frequencies of different voice signal types. We found that if the energy is concentrated in the harmonics, the signal is of a lower voice type. Aspiration noise and aperiodicity both contribute to a lower SCR, which explains the small SCR of type 4 signals. These results indicate that the underlying mechanism of disordered voice production may be nonlinear.

CONCLUSION

STFT was used to analyze voice signals and produce an objective parameter called SCR for quantifying the aperiodicity of voice signals. This continuous variable is an objective parameter capable of quantifying all four types of voice signals.

Since SCR can be calculated quickly, it offers a quick evaluation of a clinical subject's voice. By supplementing SCR with previous analysis tools, clinicians get more general acoustic information about a subject and can prescribe corresponding treatments. Future work with SCR could establish the criterion for all 4 types and provide concrete boundaries between voice signals. By controlling the ratio of vibration and turbulence portions of voice signals, future studies could synthesize type 1, 2, 3 and 4 voice signal samples to allow observation of the relationship of the SCR to voice signal and establish objective boundaries between each type. Additionally, by following how SCR changes during specific treatment interventions, treatment methods could be evaluated for efficacy.

ACKNOWLEDGMENTS

This study was supported by NIH grant number R01-DC006019 from the National Institutes of Health.

REFERENCES

1. Titze, IR. Notional Center for Voice and Speech, CO; Denver.: 1995. Workshop on acoustic voice analysis: Summary statement; p. 26-27.
2. Sprecher A, Olszewski A, Jiang JJ, Zhang Y. Updating signal typing in voice: Addition of type 4 signals. *The Journal of the Acoustical Society of America*. 2010; 127(6):3710–3716. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20550269>. [PubMed: 20550269]
3. Titze, IR.; Baken, R.; Herzel, H. *Vocal Fold Physiology: Frontiers in Basic Science*. San Diego; Singular, CA: 1993. Evidence of Chaos in Vocal Fold vibration; p. 143-188.
4. Jiang JJ, Zhang Y, McGilligan C. Chaos in voice, from modeling to measurement. *Journal of Voice*. Mar; 2006 20(1):2–17. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15964740>. [PubMed: 15964740]
5. Jiang JJ, Zhang Y, Ford CN. Nonlinear dynamic of phonations in excised larynx experiments. *The Journal of the Acoustical Society of America*. Oct; 2003 114(4 Pt 1):2198–2205. Available: <http://www.ncbi.nlm.nih.gov/pubmed/14587617>. [PubMed: 14587617]
6. Little MA, MacSharry PE, Roberts SJ, Costello DAE, Moroz IM. Exploiting Non-linear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *BioMedical Engineering [Online]*. Jun.2007 6(23) Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1913514/>.

7. Allen JB. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Transactions on Acoustics, Speech, Signal Processing. Jun.1977 ASSP-25:235–238.
8. Allen JB. Application of the short-time Fourier transform to speech processing and spectral analysis. Proc. IEEE ICASSP-82. :1012–1015, 1982.
9. Cohen, L. Time-Frequency Analysis. Prentice-Hall; Englewood Cliffs, NJ: 1995.
10. Selesnick W, Baraniuk RG, Kingsbury N. The dual-tree complex wavelet transform - A coherent framework for multiscale signal and image processing. IEEE Signal Processing Magazine. Nov; 2005 22(6):123–151. Available: <https://scholarship.rice.edu/bitstream/handle/1911/20355/Sel2005Nov1TheDualTre.PDF?sequence=1>.
11. Packard NH, Crutchfield JP, Farmer JD, Shaw RS. Geometry from a Time Series. Phys. Rev. Let. Sep; 1980 45(9):712–716. Available: <http://journals.aps.org/prl/abstract/10.1103/PhysRevLet- t.45.712>.
12. Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. Physical Review. Feb; 1986 33(2):1134–1140. Available: <http://chaos.utexas.edu/manuscripts/1064949034.pdf>. [PubMed: 9896728]
13. Theiler J. Spurious dimension from correlation algorithms applied to limited time-series data. Physical Review A. Sep; 1986 34(3):2427–2432. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9897530>.
14. Grassberger P, Procaccia I. Characterization of Strange Attractors. Physical Review Letters. Jan; 1983 50(5):346–349. Available: <http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.50.346>.
15. Grassberger P, Procaccia I. Characterization of strange attractors. Phys Rev Lett. 1983; 50:346–349.
16. Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Phys D. 1983; 9:189–208.

d)

Fig 1.

A) Convergence graph of a type 1 voice signal. B) Spectrogram of a type 1 voice signal. C) Convergence graph of type 1 voice signal. D) Spectrogram of a type 4 voice signal.

Fig 2.

A) Box plot of SCR at PPS 0.012 for all four voice signal types.

Fig 3.

A) Box plot of shimmer for type 1 and type 2 voice signals. B) Box plot of jitter for type 1 and 2 voice signals C) Box plot of Correlation Dimension for type 2 and 3 voice signals.

Fig 4.
Histogram distribution plots of SCR distribution for all four voice types at PPS 0.012.

Fig 5.

A) Performance comparison between SCR and shimmer in voice classification. B) Performance comparison of SCR and jitter in voice classification. C) Performance comparison between SCR and Correlation Dimension in voice classification.

TABLE I

Subject Characteristics

Voice Type	Number of Samples	Age in years
<i>Type 1</i>	22	37.3 (22-63)
<i>Type 2</i>	49	44.5 (23-75)
<i>Type 3</i>	51	49.9 (7-80)
<i>Type 4</i>	26	62.8 (30-85)

Subject characteristics for each signal type. Age is displayed as mean age (age range)

TABLE II

SCR Voice Comparison PPS = 0.012

Comparison	P
<i>Type 1 vs Type 2</i>	<.001
<i>Type 1 vs Type 3</i>	<.001
<i>Type 2 vs Type 3</i>	<.001
<i>Type 3 vs Type 4</i>	<.001

Pairwise t-tests examining if SCR was successful at differentiating between each of the 4 voice types at PPS 0.012.

TABLE III

Comparison	P
<i>Type 1 vs Type 2</i>	.423
<i>Type 1 vs Type 3</i>	.010
<i>Type 2 vs Type 3</i>	0.02

Pairwise t-tests for correlation dimension analysis as a tool to differentiate between voice signal type. D2 was unable to differentiate between type 1 and type 2 signals or quantify type 4 voice signals.