The Effect of Moving Window on Acoustic Analysis

*Min Shu, *,†Jack J. Jiang, and †Malachi Willey, *Shanghai, China, and †Madison, Wisconsin

Summary: Objective. To investigate the effects of the moving window method on acoustic measures and discrimination ability between normal and disordered voices.

Methods. Fifty-three normal voices and 50 disordered voices were recruited. Three selection methods, the moving window method, the mid-vowel method, and the whole vowel method, were applied to each raw audio signal to determine the most stable segment of each signal. Acoustic parameters such as percent jitter, percent shimmer, signal-to-noise ratio (SNR), cepstral peak prominence (CPP), and correlation dimension (D2) were calculated. The Wilcoxon test was used to compare the stability of these segments across different methods. An artificial neural network was used for estimating how well disordered voices were discriminated from normal ones.

Results. Segments selected using the moving window method were more stable than those selected using the other two methods, meaning lower perturbation and nonlinear dynamic measurements as well as higher SNR and CPP values. The discrimination accuracy rate for the moving window method was 91.90 ± 8.73%, whereas the mid-vowel method and the whole vowel method were 72.34 ± 12.94% and 70.34 ± 5.24%, respectively.

Conclusion. The moving window method is capable of providing a more stable audio segment and can discriminate disordered voices from normal ones more effectively.

Key Words: Moving window–Acoustic analysis–Artificial neural network.

INTRODUCTION

Acoustic analysis is an objective method for assessing the characteristics of audio signals, which has been of paramount importance in voice research. It aids in the understanding of normal and pathological voices, prompts further exploration of the phonation mechanism, and provides objective evidence for treatment efficacy evaluation.

Currently, most acoustic analysis methods are based on frequency extractions. Stable waves with small frequency fluctuations help provide accurate and reliable measures. Titze[1] also pointed out that objective assessment should be used for periodic or nearly periodic signals. Therefore, determining how to select a stable signal segment is the first step for reliable acoustic assessments.

Sustained vowels are commonly used for acoustic analysis in current research. The most widely used methods for selecting audio segments include subjectively selected vowels, mid-vowels, and whole vowels excluding onset and offset. Subjectively selected vowels are determined visually to be the most stable portion of an audio signal.[2,3] Selecting segments with this method is very subjective and may cause significant differences between raters. Although the amplitude of voices may be intuitively judged by the naked eye, it is difficult to judge stability based on characteristics such as frequency and signal-to-noise ratio (SNR). Middle vowels are captured equivalently on either side of the exact mid-point of audio signals.[4–6] Whole vowels refer to segments which exclude voice onset and offset. The periods of onset and offset range from 200 ms to 1 s in literature.[7,8] These two methods are applied based on the same consideration that the onset and offset of a signal lack stability due to the interaction between the laryngeal muscles and vocal folds. This interaction causes changes in glottal aerodynamic and biomechanical properties.[2,6,9] Therefore, the middle portion of a raw audio signal is believed to be the most stable segment.

However, previous studies have shown that several subjective factors might alter the acoustic analysis results of whole vowels and middle vowels, making the results incomparable across laboratories. One issue is that each person has different onset and offset times. Hoit et al[10] reported that voice onset time is generally longer at high lung volumes. Higgins et al[11]also noted that alterations of glottic shape caused by external forces would influence the onset time. Due to these factors, applying a standardized extraction rule on different individuals may result in an unreliable and incomparable acoustic assessment. Furthermore, no signal is absolutely stable. Even in normal voices, small fluctuations are observed in frequency or amplitude.[4,12] Therefore, acoustic analysis results may vary with the positions of the selected points. A stable segment can be obtained at any position of the entire signal, including the beginning and middle points.

The moving window method refers to the process in which a window of a certain length moves along the sample incrementally and provides sequential segments. Segments of interest will be selected and analyzed. This method is usually used in the analysis of electroencephalogram (EEG),[13] electrocardiogram (ECG),[14,15] electromyogram (EMG),[16] and genetics studies. In this study, a *MATLAB* (MathWorks, R2012b) program was developed to automatically cut the raw signal into sequential segments (Figure 1). Acoustic measures of the segments selected using the moving window method will be compared with those selected by the mid-vowel and whole vowel methods. This comparison will determine which method can provide the most stable segment.

Many biological and physiopathological processes have "chaotic" features and cannot be accurately analyzed by conventional linear analysis methods. An artificial neural network

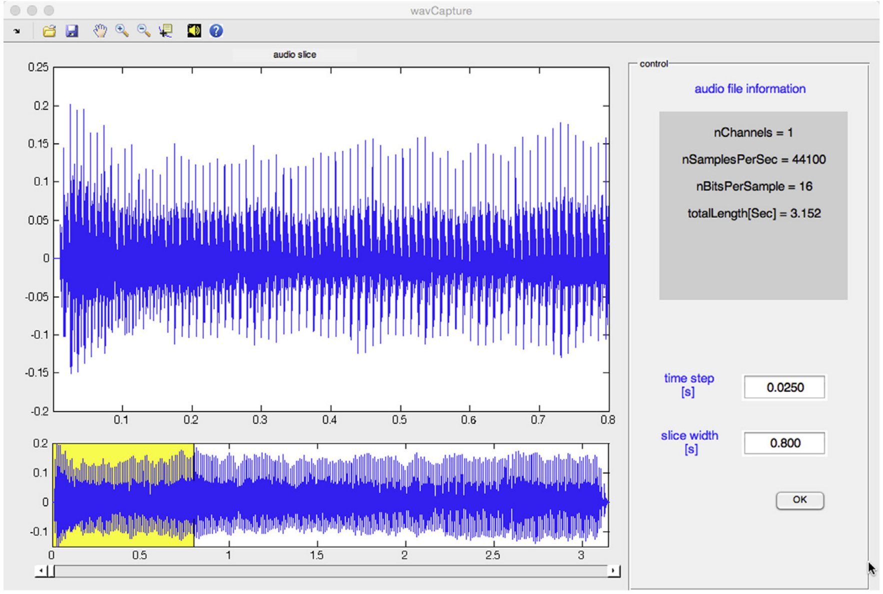
Accepted for publication November 20, 2014.

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

Address correspondence and reprint requests to Jack J. Jiang, Department of Otolaryngology Head & Neck Surgery, Eye, Ear, Nose and Throat Hospital, Fudan University, 83 Fenyang Road, Shanghai 200031, China. E-mail: jjjiang@wisc.edu

Journal of Voice, Vol. 30, No. 1, pp. 5-10
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.11.008



wavCapture

audio slice

control

audio file information

nChannels = 1

nSamplesPerSec = 44100

nBitsPerSample = 16

totalLength[Sec] = 3.152

time step [s] 0.0250

slice width [s] 0.800

OK

FIGURE 1. User interface of a *MATLAB* (MathWorks, R2012b) program developed to sequentially partition the raw audio signal into segments of a defined length. An 800 ms window (the yellow rectangle in the bottom window on the left) moves along the whole audio signal (the bottom window on the left) from the beginning (onset) to the end (offset) with a 25 ms time shift. The top window shows the 800 ms selected segment as a yellow rectangle. (For interpretation of references to color in this figure legend, the reader is referred to the web version of this article.)

(ANN) is a mathematical model that simulates the structure and function of biological neural networks with large scale parallel processing features. It is a nonlinear statistical data modeling tool used to model complex relationships between inputs and outputs. The nonlinear analysis features make it suitable for many medical applications. An early application of neural networks was for the diagnosis of acute myocardial infarction patients. The diagnostic sensitivity of the neural network was found to be from 92 % to 97 % with a specificity of 95–96%. At the same time, the sensitivity and specificity of conventional diagnostic methods were only 78–88% and 74–85%, respectively.[17–19] In this study, an ANN will be used to determine whether the moving window method can more effectively differentiate pathological from normal voices when compared with the other two methods.

MATERIALS AND METHODS

Subjects

Fifty-three normal voices (21 males and 32 females) were recruited with an age range of 22–59 years and an average age of 36.0. Fifty disordered voices (19 males and 31 females) were included with an age range of 21–65 years and an average age of 43.1. Table 1 summarizes the variety of voice disorders included in the sample. All voice samples were selected from KayPentax Disordered Voice Database (Model 4337, Version 1.03, Kay Elemetrics Corp, Lincoln Park, NJ, developed by Massachusetts Eye and Ear Infirmary Voice and Speech Lab).

Vowel selection

Each sample was selected in three different methods:

Mid-vowel: A 250 ms segment was selected from both sides of the midpoint of an utterance of the prolonged vowel. The 500 ms segment was used for further analysis.

Whole vowel: An entire utterance of the prolonged vowel minus 200 ms from the beginning and 200 ms from the end was used for acoustic analysis. This process excluded the onset and offset segments that are believed to be the most instable parts of an audio signal.

Moving window: An 800 ms window length and 25 ms time shift was applied in the current study. A range of window lengths (100–1000 ms) and time shifts (25–100 ms) were evaluated by comparing perturbation and other measures using the Wilcoxon test. A window length of 800 ms and shift time of 25 ms were determined to be optimal values to avoid large variability within a small segment and variability due to subject effort differences across tokens. Olszewski et al[20] also suggested that window lengths between 550 and 950 ms provided the best results in both segment length and stability. Each raw signal was cut into sequential segments using this method. Acoustic measures of each sequential segment were normalized and summed. The segment with the highest score was

TABLE 1.
Distribution and Number of Recruited Pathological Voices

Pathological Type	Number
Contact granuloma	2
Gastric reflux	16
Paralysis	7
Scarring	3
Edema	12
Polyp	4
Nodules	6
Total	50