American Journal of Biomedical Engineering & Technology

American Journal of
Biomedical Engineering & Technology
http://ivyunion.org/index.php/ajet
Vol. 2, Article ID 201400622, 13 pages

Research Article

Exploiting Wavelet and Prosody-related Features for the Detection of Voice Disorders

Upal Mahbub[1] and Celia Shahnaz[2*]

[1] Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland, USA

[2] Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Abstract

An approach for the detection of voice disorders exploiting wavelet and prosody-related properties of speech is presented in this paper. Based on the normalized energy contents of the Discrete Wavelet Transform (DWT) coefficients over all voice frames, several statistical measures are first determined. Then, the idea of some prosody-related voice properties, such as mean pitch, jitter and shimmer are utilized to compute similar statistical measures over all the frames. A set of statistical measures of the normalized energy contents of the DWT coefficients is combined with a set of statistical measures of the extracted prosody-related voice properties in order to form a feature vector to be used in both training and testing phases. Two categories of voice samples namely, healthy and disordered are considered here thus formulating the problem in the proposed method as a two-class problem to be solved. Finally, an Euclidean Distance based classifier is used to handle the feature vector for the purpose of detecting the disordered voice. A number of simulations is carried out and it is shown that the statistical analysis based on wavelet and prosody-related properties can effectively detect a variety of voice disorders from the mixture of healthy and disordered voices.

Keywords: Voice Disorder; Wavelet Transform; Pitch; Jitter; Shimmer; Statistical Measures

Academic Editor: Sihua Peng, PhD, Shanghai Ocean University, China

Received: February 11, 2015; **Accepted**: June 5, 2015; **Published:** June 20, 2015

Competing Interests: The authors have declared that no competing interests exist.

***Correspondence to**: Celia Shahnaz, Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Bangladesh; **Email**: celia@eee.buet.ac.bd

1. Introduction

Any pathological condition which distorts or hampers the normal flow of voice may be considered as voice disorder [1] - [4]. Investigation revealed that voice samples carry symptoms of disorder in the place of their origin. The common disorders are acute infective laryngitis, chronic non-specific laryngitis, vocal fold paralysis etc. Physicians often use invasive techniques like Endoscopy to diagnose the symptoms of vocal fold disorders. However, it is possible to identify disorders using certain prosody-related properties of voice Signals [1].

In disordered voice, the non-stationary behavior of the voice can be analyzed in time, frequency and time-frequency domains [5]. The energy distribution in various levels of scaling of Wavelet Transforms can provide information about localized irregularity in vocal fold vibration. In the event of paralysis, voiced speech samples contain irregular pitch due to improper functioning of vocal cords. Jitter is the periodic fluctuation of the pitch. However, these fluctuations were found to be more erratic for pathological voices such as functional voice disorders and pathologic larynges. Shimmer is the cycle-to-cycle variability of the period amplitude of vocal fold vibration. It is used as one of the measures of micro-instability of the vocal fold vibrations.

This paper uses wavelet analysis technique and prosody-related voice properties to form the basic feature vector, which is used as input to an Euclidean distance based classifier. Statistical operations, such as, mean, standard deviation, median, range, minimum value, maximum value etc. on the normalized energy contents of the Discrete Wavelet Transform (DWT) coefficients are found to be useful while formulating a feature vector for identifying pathological disorders in the larynx. It is shown by further analysis that, inclusion of similar statistical measures of the prosody related voice properties, such as, mean pitch, jitter and shimmer in the feature vector are important and significant in providing a better discrimination between healthy and disordered voice samples.

2. Proposed Method

The block diagram of the training/testing phase of the proposed method is shown in Fig. 1. The work-flow includes collection of speech samples of healthy (H) and voice disordered (VD) persons for training purpose. The feature extraction procedure includes wavelet analysis, and calculation of some prosody-related speech properties, namely, pitch, jitter and shimmer. On performing some statistical analysis on the wavelet coefficients, and that on the extracted prosody-related speech properties, a feature vector is formed to classify the test data in one of the two groups, H or VD, based on a Euclidean distance based classifier. In this Section, first, some insight of the theoretical aspects of wavelet transform, pitch, jitter and shimmer are presented. Then, the formulation of the feature vector based on statistical analysis is described.

2.1 Wavelet Analysis

Wavelet transform is an effective tool for analysis of non-stationary signals, such as speech, since it is more useful in localizing a symptom both in time and frequency scales compared to Fourier Transform. The basis functions of wavelet transform are scaled and shifted versions of the time-localized mother wavelet. Wavelet transform is an ideal tool for determining whether or not a signal is stationary in a global sense. When a signal is judged non-stationary, the wavelet transform can be used to identify stationary sections of the data stream.

The Discrete Wavelet Transform (DWT), which is based on sub-band coding, is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. The Discrete Wavelet Transform (DWT) involves choosing scales and positions based on powers of two, so called dyadic scales and positions. The mother wavelet is rescaled or dilated by powers of two and translated by integers. For a speech signal $x(n)$, a general equation for the discrete wavelet transform (DWT) transformed signal can be written as [6],

$$X[a,b] = \sum_{-\infty}^{\infty} x[n] \frac{1}{\sqrt{a}} \Psi^{*}[\frac{n-b}{a}], \tag{1}$$

where, b is a real number known as window translation parameter and a is a positive real number named as dilation or contraction parameter, $*$ denotes complex conjugate and Ψ^{*} represents the wavelet function. In case of DWT, the signal is passed through a series of highpass and lowpass filters to decompose it into approximate and detail information/coefficients thus analyzing the signal at low as well as high frequency bands with different resolution [6]. Fig. 2 shows the wavelet decomposition for DWT with a scale of 3 levels, where, filters of different cut-off frequencies are used for analyzing the signal at different scales. Filtering operations change the resolution of the signal, whereas sub sampling (down/up sampling) results in the change of the scale. Thus, DWT decomposes the signal into approximate and detail information thereby helping in analyzing it at different frequency bands with different resolutions. In Fig. 2, down arrow $(\downarrow)$ represents down sampling. At each level of the figure the signal is decomposed into low and high frequencies. Due to the decomposition process the input signal must be a multiple of 2^{n} where n is the number of levels. For example a signal with 32 samples, frequency range 0 to f_n and 3 levels of decomposition, 4 output scales are produced as shown in Table 1.

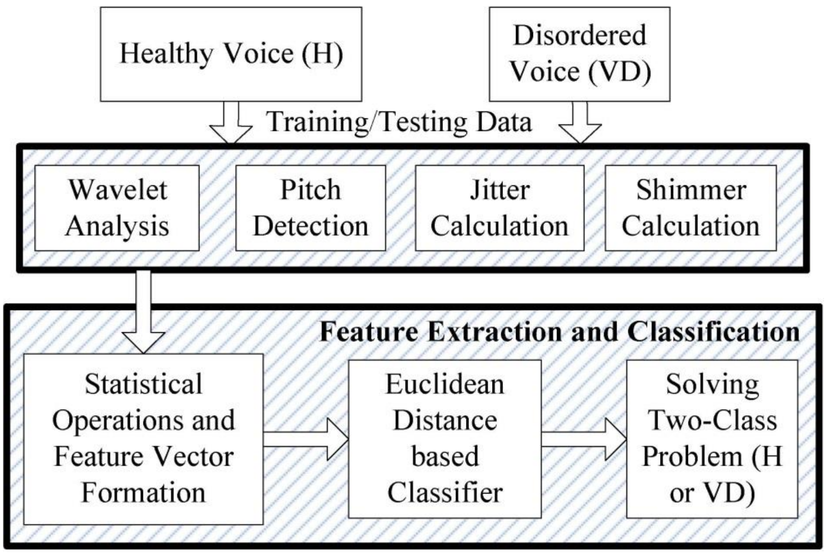

Fig. 1 Block diagram of training/testing phase of the proposed method.

Table 1
Three level discrete wavelet decomposition

Level	Frequencies	Samples
3	0 to $\frac{f_n}{8}$	4
	$\frac{f_n}{8}$ to $\frac{f_n}{4}$	4
2	$\frac{f_n}{4}$ to $\frac{f_n}{4}$	8
1	$\frac{f_n}{8}$ to f_n	16

A Filter Bank is used to extract the wavelet coefficients. Wavelet packet decomposition at level three was applied to the speech signal using the db2 wavelet filters. The lower band scale presents a more dominant periodicity than the higher band scale. This periodicity is decreasing in the pathological speech but it is very consistent in the normal speech. In Table 2, the decomposition and reconstruction filter coefficients for db2 wavelet is shown.

2.2 Prosody-related Speech Features

For normal speech, the vocal fold would vibrate at regular intervals with a fundamental frequency F_0 termed as pitch which is a prosodic property of speech. Pitch is an important property of voiced speech signal. Among several techniques for pitch detection [7], in this work, we have employed the commonly used autocorrelation method. It is based on the idea that if correlation is done on a speech signal with itself, then the period of the autocorrelation function exhibits the period of the fundamental frequency of speech. The value of pitch may vary from time to time and this variation is known as jitter. Therefore, jitter, a prosody related property, can be calculated as,

$$Jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|P_i - P_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}P_i}, \quad (2)$$

where, P_i is the value of pitch at i-th frame and N is the total number of frames under consideration.

As the human phonatory system is not a perfect machine, jitter is always present even in normal speech [8].

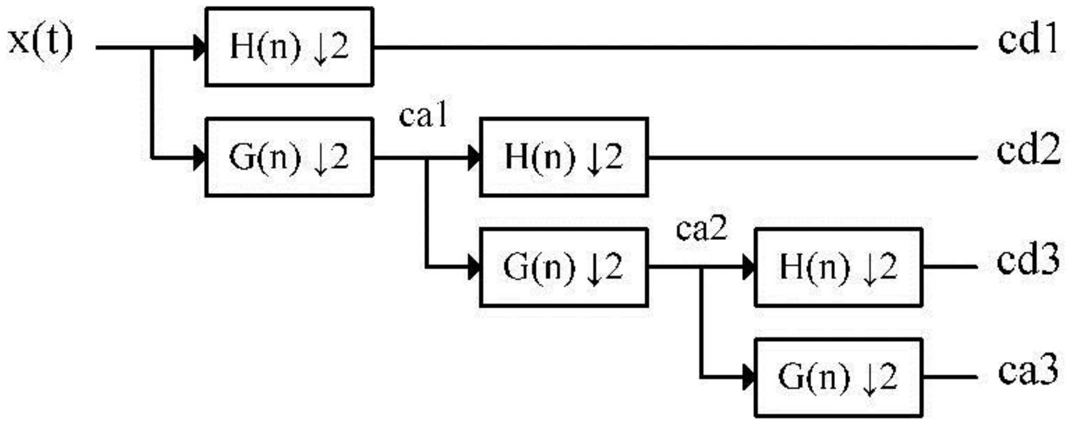

Fig. 2 Three level Discrete Wavelet Decomposition

Recent researches have shown that glottal waveform dynamics can play an important role in voice identification [8]. In the context of glottal waveform, shimmer is another prosody related property that represents the cycle-to-cycle variability of the period amplitude of vocal fold vibration at glottal opening or closing instances. Thus, for the calculation of shimmer, it is necessary to detect of glottal opening or closing instances and the amplitude of speech at those instances. Shimmer, usually expressed in dB, can be computed as,

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1} 20log(\frac{A_{i+1}}{A_i}), \quad (3)$$

where, A_i is the signal amplitude at glottal opening/closing instance.

A flowchart for the detection of the prosody related properties extracted from the speech signal, such as, mean pitch over several cycles of a frame, jitter and shimmer, is shown in Fig. 3.

Table 2

Daubechies-2 tap (db2) wavelet coefficients

Decomposition Filter Coefficients		Reconstruction Filter Coefficients	
LPF	HPF	LPF	HPF
−0.1294	−0.4829	0.4829	−0.1294
0.2241	0.8365	0.8365	−0.2241
0.8365	−0.2241	0.2241	0.8365
0.4829	−0.1294	−0.1294	−0.4829

2.3 Feature Extraction based on Statistical Analysis

In the training phase, at every frame, we intend to calculate the energy contents of the speech at various levels (scaling factor) of the DWT coefficients. For three level decomposition resulting in five DWT coefficients, the energy content of each DWT coefficient is normalized against total energy content in the signal. The normalized energy across the scale i is,

$$E_N(i) = \frac{E_i}{E_T} \tag{4}$$

where $i = 1,2,3...$ corresponds to different levels of DWT, E_T represents the total energy across all the levels and E_i stands for the energy at i-th level. The statistical measures, such as, average (AVG), median (MED), standard deviation (STD), minimum (MIN), maximum (MAX), range (RNG) and interquartile range (IQR) of the normalized energy contents of the DWT coefficients are obtained over all the frames of a speech sequence. The working formula for these statistical measures are summarized in Table 3. Similarly, as for prosody related properties, i.e. for mean pitch, jitter and shimmer, average and standard deviation over all the frames are determined. Finally, the statistical measures of the normalized energy contents of the DWT coefficients and those of the prosody related speech properties extracted over all the speech frames are exploited to form a feature vector for both training and testing phases.

In addition to considering all the statistical measures at the same time, we propose to employ several combinations of the statistical measures obtained from the DWT and prosody in order to reduce the feature dimension and to find the desired feature vector that is effective in identifying voice disorder. In the proposed feature vector, the combination of the statistical measures includes (i) only the AVG of the normalized energy contents of the DWT coefficients (AVG-W), (ii) only the STD of the normalized energy contents of the DWT coefficients (STD-W), (iii) all the statistical measures (as mentioned in Table 3) of the normalized energy contents of the DWT coefficients (Stat-W), (iv) only the median of the normalized energy contents of the DWT coefficients (MED-W) (v) the minimum values of the normalized energy contents of the DWT coefficients (MIN-W) (vi) the maximum values of the normalized energy contents of the DWT coefficients (MAX-W) (vii) only the range of the MAX and MIN values of the normalized energy contents of the DWT coefficients (RNG-W) (viii) the Interquartile Range of the normalized energy contents of the DWT coefficients (IQR-W) (ix) STD-W along with the STD of mean pitch, values of jitter and shimmer (STD-WPJS), (x) Stat-W as well as the STD of mean pitch, and the values of jitter and shimmer (Stat-WPJS) (xi) STD-W and the STD of mean pitch, and the STD of the signal amplitude at glottal opening instances (STD-WPAmpl), (xii) STD-W and the STD of the signal amplitude at glottal opening instances (STD-WAmpl), (xiii) STD-W and the value of jitter, and the STD of the signal amplitude at glottal opening instances (STD-WJAmpl). In the testing phase, similar statistical analysis is performed on the test speech to form a feature vector.

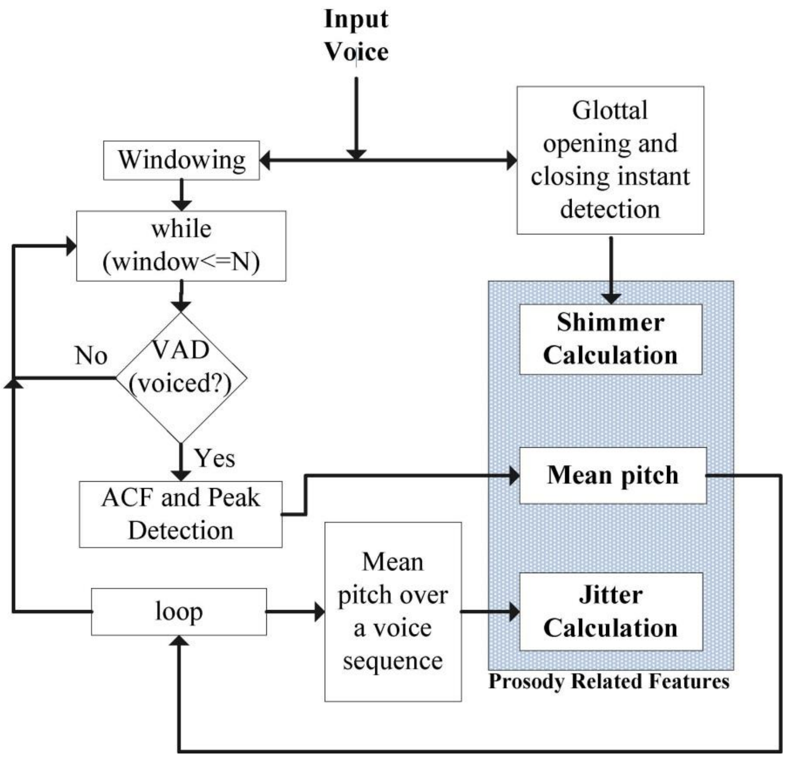

Fig. 3 Detection of mean pitch over several cycles of a frame, jitter and shimmer from speech

Table 3
Basic statistical measures

Statistic	Equation
Average (AVG)	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
Median (MED)	50 -th percentile
Standard Deviation (STD)	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
Minimum (MIN)	5 -th percentile
Maximum (MAX)	95 -th percentile
Range (RNG)	$MAX - MIN$
Interquartile Range (IQR)	75 -th percentile - 25 -th percentile

2.4 Voice Disorder Detection

In the proposed method, for the purpose of voice disorder detection using the extracted feature vector, a distance based similarity measure is utilized [1]. The detection task is carried out based on the distances of the feature vectors of the training speech sequences from the feature vector of the testing speech sequence. Given the m-dimensional feature vector for the k-th training speech sequence of the j-th speaker be $\{\gamma_{jk}(1), \gamma_{jk}(2), \ldots\ldots, \gamma_{jk}(m)\}$ and a f-th test speech sequence with a feature vector $\{\upsilon_f(1), \upsilon_f(2), \ldots\ldots, \upsilon_f(m)\}$, a similarity measure between the test speech sequence f of the unknown speaker and the training speech sequences of the j-th speaker is defined as,

$$D_j^f = \sum_{k=1}^{q}\sum_{i=1}^{m} |\gamma_{jk}(i) - \upsilon_f(i)|^2, \tag{5}$$

where, a particular class represents a speaker with q number of training speech sequences. Therefore, according to (5), given the f-th test speech sequence, the unknown speaker is identified as the speaker j among the p number of classes, when,

$$D_j^f < D_g^f, \forall j \neq g, and, \forall g \in 1,2,\ldots, p. \tag{6}$$

In our case, we are interested to handle a two-class problem $(p = 2)$, i.e. to discriminate between healthy (H) and voice disordered (VD) speakers.

3. Simulation Results

A number of simulations is performed to evaluate the efficacy of the proposed method in terms of standard metrics, such as sensitivity, specificity and accuracy. The training group consists of 10 independent speech sequences from 10 different speakers including both male and female having different types of speech disorders, namely unilateral and bilateral vocal fold paralysis, presbylaryngis caused by thinning of vocal fold muscles and tissues with aging, spasmodic dysphonia (SD) caused by involuntary movement of one or more muscles of the larynx etc. On the other hand, we consider 5 speech sequences, including both male and female, of speakers having no voice disorder. Each of the training speech sequences are of 9.375 seconds duration with sampling frequency of 16 kHz and they contains the same text uttered by different speakers. The disordered speech sequences are obtained from the online samples of ‘Texas Voice Center’ [9]. While determining the feature vector in the testing phase, in order to detect pitch, voiced speech segments are used, where the periodicity of the vocal fold opening and closing is clearly observed.

For testing purpose, 10 disordered speech sequences and 5 normal speech sequences, including both male and female speakers, are taken into account. Although the speakers in the testing phase are the same as those in the training phase, none of the speech sequences used in the testing are employed in the training. Unlike the results reported in [10], in order to evaluate the effect of length of speech sequence on the detection performance, the test speech sequences are arranged in five different groups. In Group 1, long sequences of 8.125 seconds each are adopted from each of the H and VD speakers. In Group 2, a set of sequences, where two smaller sequences of equal length of 4.0625 seconds obtained by diving each of the long sequences are considered. The same long sequences are divided into three sequences of equal length thus forming a new set of sequences in Group 3. In Group 4, long sequences of 8.125 seconds each are adopted from five of the H and five VD speakers and finally in Group 5, 8.125 second sequences are adopted from five of the H and four VD speakers. Thus Group 1 consists of 15 (10 disordered and 5 normal), Group 2 contains 30 (20 disordered and 10 normal), Group 3 possesses 45 (30 disordered and 15 normal), Group 4 consists of 10 (5 disordered and 5 normal) and Group 5 contains 9 (4 disordered and 5 normal) test speech sequences. The classification results of Group 1, 2, 3, 4 and 5 for different sets of feature vectors are shown in Tables 4 through 8, sequentially.

In the analysis of simulation results, sensitivity reflects the rate at which disorder is detected given that the test speaker truly has been diagnosed with it. Specificity means the rate at which disorder is correctly ruled out as being present when the test speaker truly has no diagnosis. Accuracy is computed simply as the total number of correct classifications divided by the total number of possible observations. The performance of the proposed method is evaluated for each of the combinations of the statistical measures (AVG-W, STD-W, Stat-W, MED-W, MIN-W, MAX-W, RNG-W, IQR-W, STD-WPJS, Stat-WPJS, STD-WPAmpl, STD-WAmpl, STD-WJAmpl) forming the feature vector by computing the resulting sensitivity, specificity, and accuracy as follows [11],

$$Sensitivity = \frac{TP}{TP+FP}, \tag{7}$$

$$Specificity = \frac{TN}{TN+FN}, \tag{8}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}, \tag{9}$$

where, TP represents true positives (i.e., correct classification as disordered), FP stands for false positives (i.e., incorrect classification of test speech sequence as disordered), TN symbolizes true negatives (i.e., correct classification as not disordered), and FN specifies false negatives (i.e., incorrect classification of disordered as not disordered).

Table 4
Classification results - for group 1

Feature Group	Sensitivity	Specificity	Accuracy
AVG-W	1	0.7143	0.8667
STD-W	1	0.7143	0.8667
Stat-W	0.9000	0.8000	0.8667
MED-W	0.6923	0.5000	0.6667
MIN-W	1	0.8333	0.9333
MAX-W	0.8182	0.7500	0.8000
RNG-W	0.8889	0.6667	0.8000
IQR-W	0.9091	1	0.9333
STD-WPJS	0.7273	0.5000	0.6667
Stat-WPJS	0.7273	0.5000	0.6667
STD-WPAmpl	0.6667	0.3333	0.5333
STD-WAmpl	1	1	1
STD-WJAmpl	1	1	1

Table 5
Classification results - for group 2

Feature Group	Sensitivity	Specificity	Accuracy
AVG-W	0.7500	0.5000	0.6667
STD-W	0.8824	0.6154	0.7667
Stat-W	0.7895	0.5455	0.70
MED-W	0.6538	0.2500	0.6000
MIN-W	0.9048	0.8889	0.9000
MAX-W	0.7778	0.5000	0.6667
RNG-W	0.7778	0.5000	0.6667
IQR-W	0.7917	0.8333	0.8000
STD-WPJS	0.7895	0.5455	0.7000
Stat-WPJS	0.7895	0.5455	0.7000
STD-WPAmpl	0.7059	0.3846	0.5667
STD-WAmpl	0.9500	0.9000	0.9333
STD-WJAmpl	0.9474	0.8182	0.9000

Table 6
Classification results - for group 3

Feature Group	Sensitivity	Specificity	Accuracy
AVG-W	0.6786	0.3529	0.5556
STD-W	0.8276	0.6250	0.7556
Stat-W	0.7586	0.5000	0.6667
MED-W	0.7333	0.4667	0.6444
MIN-W	0.9333	0.8667	0.9111
MAX-W	0.7500	0.4706	0.6444
RNG-W	0.7500	0.4706	0.6444
IQR-W	0.7059	0.4545	0.6444
STD-WPJS	0.7143	0.4118	0.6000
Stat-WPJS	0.7143	0.4118	0.6000
STD-WPAmpl	0.8077	0.5263	0.6889
STD-WAmpl	0.8929	0.7059	0.8222
STD-WJAmpl	0.9231	0.6842	0.8222

Table 7
Classification results - for group 4

Feature Group	Sensitivity	Specificity	Accuracy
AVG-W	1.0000	0.7143	0.8000
STD-W	1.0000	0.7143	0.8000
Stat-W	0.7500	0.6667	0.7000
MED-W	0.6000	0.6000	0.6000
MIN-W	0.8000	0.8000	0.8000
MAX-W	0.6000	0.6000	0.6000
RNG-W	0.6000	0.6000	0.6000
IQR-W	0.8000	0.8000	0.8000
STD-WPJS	0.5000	0.5000	0.5000
Stat-WPJS	0.5000	0.5000	0.5000
STD-WPAmpl	0.5714	0.6667	0.6000
STD-WAmpl	1	1	1
STD-WJAmpl	1	1	1

Table 8
Classification results - for group 5

Feature Group	Sensitivity	Specificity	Accuracy
AVG-W	1.0000	0.7143	0.7778
STD-W	1.0000	0.7143	0.7778
Stat-W	0.6667	0.6667	0.6667
MED-W	0.5000	0.6000	0.5556
MIN-W	0.8000	1.0000	0.8889
MAX-W	0.5000	0.6000	0.5556
RNG-W	0.5000	0.6000	0.5556
IQR-W	0.7500	0.8000	0.7778
STD-WPJS	0.4000	0.5000	0.4444
Stat-WPJS	0.4000	0.5000	0.4444
STD-WPAmpl	0.5000	0.6667	0.5556
STD-WAmpl	1	1	1
STD-WJAmpl	1	1	1

It can be seen from Tables 4 through 8 that employing the feature vector that is a combination of standard deviation of normalized energy contents of the DWT coefficients and the standard deviation of amplitude at glottal opening instances (STD-WAmpl), yields the best results in the sense of higher sensitivity, specificity and accuracy irrespective of the length of the test speech sequences and the ratio of number of speech samples at the two separate classes. It can also be observed that the proposed technique of disorder detection performs better for larger data sequences in comparison to the scenario using smaller sequences.

4. Conclusion

This paper presents a technique for voice disorder detection utilizing the normalized energy contents of three level DWT coefficients and some prosody-related speech properties, such as, mean pitch, jitter and shimmer. It is shown that the statistical analysis on the normalized energy contents of the DWT coefficients and that on the prosody-related properties is potential and productive in formulating an effective feature vector for voice disorder detection. Extensive simulations demonstrate that the feature vector obtained from the wavelet and prosody can successfully solve the two class problem of disorder detection based on a simple yet accurate distance based classifier.

Acknowledgements

The authors would like to express their sincere gratitude towards the authorities of the Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology (BUET) for providing constant support throughout this research work.

References

1. D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. 1em plus 0.5em minus 0.4em NY: ***IEEE Press***, 2000
2. Titze IR. *Principle of Voice production*. 1em plus 0.5em minus 0.4em NJ: ***Prentice Hall***, 1994
3. Boyanov B and Hadjitodorov S. Acoustic analysis of pathological voices: a voice analysis system for screening of laryngeal diseases. ***IEEE Engineering in Medical and Biology***. 1997, 16:74-82
4. Mahmoudi Z, Rahati S, Ghasemi MM, Asadpour V, Tayarani H, and Rajati M. Classification of voice disorder in children with cochlear implantation and hearing aid using multiple classifier fusion. ***BioMedical Engineering***. 2011. [Online]. Available: http://www.biomedical-engineering-online.com/content/10/1/3
5. Umapathy K, Krishnan S, Parsa V, and Jamieson DG. Discrimination of pathological voices using time-frequency approach. ***IEEE Trans. Biomedical Engineering***. 2005, 52:421-430
6. Mallat S. A theory for multiresolution signal decomposition: Wavelet representation. ***IEEE Trans. Pattern Analysis and Machine Intelligmce***. 1989, 11:674-693
7. Shahnaz C, Zhu WP, and Ahmad MO. Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme. *IEEE Trans. Audio, Speech, Language Processing*. 2012, 12:322-335
8. Ozdas A, Shiavi RG, Silverman SE, Silverman MK, and Wilkes DM. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. ***IEEE Transactions On Biomedical Engineering***. 2004, 51:1530-1540
9. Texas voice centre, diseases of the larynx. [Online]. Available: http://www.texasvoicecenter.com/diseases.html
10. Shahnaz C, Fattah S, Mahbub U, Zhu WP, and Ahmad M. Detection of voice disorders based on wavelet and prosody-related properties. ***IEEE International Symposium on Circuits and Systems (ISCAS)***. 2012, 1030-1033
11. Moore E, Clements MA, Peifer JW, and Weisser L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. ***IEEE Transactions On Biomedical Engineering***. 2008, 55:96-107