VOICED -NON VOICED SPEECH SIGNAL CLASSIFICATION USING MULTIPLE FEATURE BASED SUPPORT VECTOR MACHINE

Dushyanth N D[1], Hareesh K[2], Christian Sauer[3], Mrityunjaya V[4]
[1]K S School of Engineering & Management, Bangalore, India
[2,3]Centre for Intelligent Computing, University of West London
[4]JSS Academy of Technical Education, Bangalore, India

Abstract

In this paper presents a novel method to classify voiced and non voiced speech segments based on the three features by extracted by analyzing the speech signal in time-frequency domain. The EMD method decomposes the speech signal into a set of intrinsic mode functions (IMFs) which is a set of narrow-band amplitude and frequency modulated components. The Zero Crossing Rate (ZCR) of IMF is used as a first feature. Using Discrete Wavelet Transform (DWT) speech signal decomposed into sub bands, Sub-band Energy (SBE) is calculated for a particular band is used as second feature. Short-Time Average Magnitude Difference Function ST-AMDF is taken as third feature.

In the process of classification of the voiced and Non-voiced speech segments these extracted features have been used as input to SVM (support vector machines) model, which is used for classification of Voiced and Non-voiced speech segments in the speech signal.

The classification algorithm was evaluated and the results were encouraging as the classification accuracy of voiced and unvoiced speech segments was greater than 96.15%. It is observed that the algorithm presented in this paper is capable of delivering higher performance in terms of classification accuracy than the other existing classification algorithms.

Keywords: Speech signal classification. Voiced-Non voiced, Empirical mode decomposition, Sub-band energy, AMDF, Support vector machine.

I. Introduction

Speech is the primary means of communication between humans and it is the dominancy of this medium that motivates research efforts to allow speech to become a viable human computer interaction [1]. The ability of machine to identify words or phrases of a spoken language and convert them in to machine readable format, so that speech can be given as input to machine is called Speech recognition.

With the rapid evolution of computer hardware and software, speech recognition has gained considerable interest in many fields like medical services, voice activated telephone exchange, industrial control, banking services, every side of society and people's lives.

The classification of speech signal into voiced, unvoiced segments provides a preliminary acoustic segmentation for speech recognition. The vibration of vocal folds produces periodic or quasi-periodic excitations to the vocal tract for voiced speech Voiced speech consists of more or less constant frequency tones of some duration. Whereas pure transient and or turbulent noises are aperiodic excitations to the vocal tract for unvoiced speech [2]. Unvoiced speech is non-periodic, random-like sounds. In recent years considerable efforts has been spent by researchers in solving the problem of classifying speech into voiced unvoiced segments [3-5].

The main aim of this classification is to extract only the voiced segments from the speech signal so that it improves the performance of speech recognition system.

The first step in any speech processing system is the extraction of features from the speech

signal. There are several parameters in speech signal; these parameters can change to different degrees in voiced-unvoiced segments. Using these parameters we can detect voiced and unvoiced region of the speech [6].

Energy is a simple measure of the loudness of the signal. This energy based method assume that speech is always louder than background noise and then assign the high-energy frames to speech and lower ones to unvoiced frames [7]. However, when the loudness of speech and noise are of similar levels for example due to the increasing background noise in the environment or during soft speech segments the simple energy feature fails to discriminate voiced and unvoiced speech segments. Noise robustness can be improved by combining energy-based features with other features, such as zero-crossing rate (ZCR) [8], or the line spectral frequency (LSF). Generally, these features work well with the speech recorded in silence or speech with high SNR value. However, when SNR falls below 10 dB or under high noise level, the discriminative power of this algorithm drops drastically.

Many earlier approaches consider the relative power of speech over the estimated noise across the different frequency bands of the spectrum to increase the discriminative power even under noisy condition. A number of researches focused on the auto-correlation of the signal to search for self-repetition components in the speech [9], such as Maximum Autocorrelation Peak, Autocorrelation Peak Count [10], which counts the number of peaks found in a range of lags. Auto-correlation-based features would fail under environments that contain repetitive noise such as motor and car noise, because the loudest frequency component is usually the noise itself.

In the most of existing methods of classification, only one feature is selected for discriminating between voiced and unvoiced speech segments. Such a method can only achieve limited accuracy because the value of any single parameter usually overlaps between categories, particularly when the speech is not recorded in a high fidelity environment. So, it becomes difficult to differentiate between voiced and non voiced speech segments using a single parameter. This leads to the motivation for more noise-robust techniques to increase the discriminating power.

The main contribution pursued in this investigation is to propose a generic optimal methodology to improve the accuracy of classification of the voiced and unvoiced speech samples. In this work 3 features are extracted by analyzing the speech signal in time-frequency domain. Intelligent models are developed using Support Vector Machines (SVM) to classify the given speech sample into voiced and unvoiced parts and this model is validated through experimental studies.

The remainder of the paper is organized as follows. Section 2 provides the details of preprocessing of speech signal. Section 3 to Section 6 explains the process of selecting several differential features with the consideration of discriminating between voiced and non-voiced speech segments.

II Speech Signal Preprocessing

If the speech signal is very much affected by the noise, preprocessing is required. Preprocessing of speech signal serves various purposes in any speech processing application. It includes noise removal, pre-emphasis, windowing and framing *etc*. **Windowing and Framing** Speech is highly variable in nature, so, rather than working on the whole signal of speech it is divided into frames and the process is termed as segmentation, which is the basic necessity of any speech processing system [11].

But, during segmentation of speech signal it might happen that the feature may split into two: half appears in one frame of the signal and the other half in another frame. The complete feature may not appear in a single analysis window and may have effectively been hidden. To overcome this problem, segmentation of speech signal is done with overlap as shown in Figure 1, to minimize the distortion.

The most widely used window for speech signal segmentation is hamming window, as it introduces least amount of distortion [11]. A typical hamming window is defined by

$$w_H = \begin{cases} 0.54 - 0.46 * \cos\left(\frac{2\pi n}{L-1}\right) & 0 \le n \le L-1 \\ 0 & otherwise \end{cases} \quad (1)$$

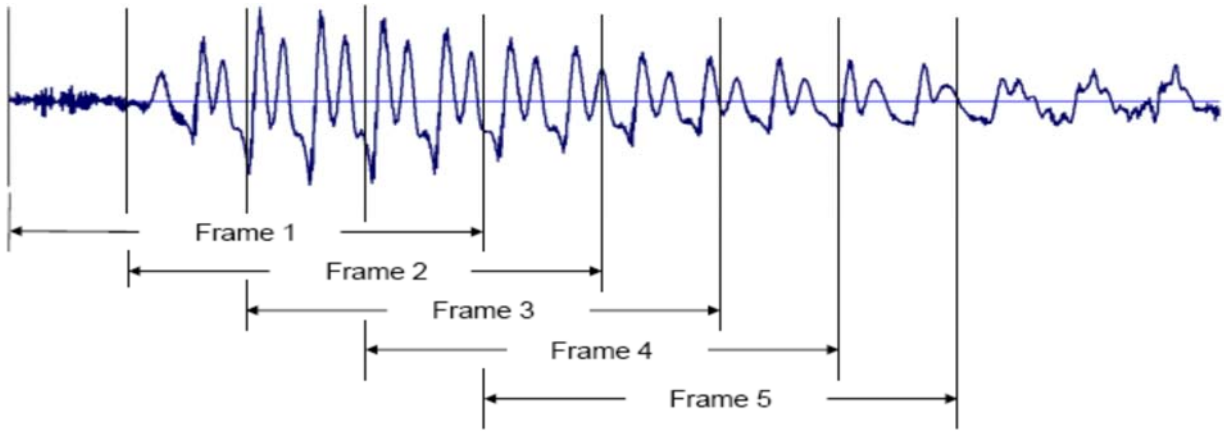

Figure 1 Speech segmentation with overlapping

Speech is processed frame-by-frame in successive windows until entire region of speech is covered by at least one such frame.
Results of analysis of individual frames are used to derive the model parameters for SVM classifier.

III. Feature Extraction

Features are the individual measurable properties of the phenomena being observed. At the foremost, speech signals are analyzed to collect features using which we can distinguish the voiced and non-voiced speech samples. The selection of features from a signal is application-specific and there is no specific procedure to select features from a signal. A signal feature which is selected should be able to.

- Clearly discriminate between voiced and non-voiced segments in speech.
- Robustness against background noise.

Choosing an independent feature for discriminating the voiced and non-voiced segment is very important for the successful classification.
In the process of signal feature selection, normally there is a trade-off between the computational time required for that feature and desired accuracy. In case of high precision and highly accurate signal feature, it may demand more computational time. So, while selecting a signal feature it is required to consider all the above mentioned aspects. In the present work, totally three features are selected by analyzing the signal both in frequency and time domain.

- Imfs Zero crossing rate
- Sub-band energy
- Short-time AMDF

IV Imfs Zero Crossing Rate

Zero-crossing rate is a crucial parameter for voiced non-voiced classification. Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a small zero-crossing count, whereas the unvoiced speech is produced by the shriveling of the vocal to cause tumultuous airflow which results in noise and shows a large zero-crossing count.

In the connection with discrete-time signals a zero crossing is said to occur if successive samples have different algebraic signs. Zero-crossing rate is a measure of number of times in a given time interval that the amplitude of the speech signals passes through a value of zero. Successive samples have different algebraic signs**,** so ZCR can be determined using equation [12].

$$\boldsymbol{ZCR} = \sum_{\boldsymbol{m}=-\infty}^{\infty} |\boldsymbol{sgn}[\boldsymbol{x}(\boldsymbol{m})] - \boldsymbol{sgn}[\boldsymbol{x}(\boldsymbol{n}-\boldsymbol{1})|\boldsymbol{w}(\boldsymbol{n}-\boldsymbol{m}) \quad \textit{(2)}$$

$$\boldsymbol{Sgn}[\boldsymbol{x}(\boldsymbol{n})] = \begin{cases} \boldsymbol{1} & \boldsymbol{x}(\boldsymbol{n}) \geq \boldsymbol{0} \\ \boldsymbol{-1} & \boldsymbol{x}(\boldsymbol{n}) < 0 \end{cases}$$

The rate at which zero crossings occur is a simple measure of the frequency content of a signal and also it is an indicator of the frequency at which the energy is concentrated in the signal spectrum.
In most of the existing method zero crossing rate is found directly on the recorded speech signal. In this work the speech signal is decomposed into a sum of the band limited functions called intrinsic mode functions

(IMFs)[15-16]. A suitable IMF is selected and its zero crossing rate is determined.

Although speech signal is non-stationary in nature, in most of the existing algorithms the Fourier transform or wavelet transform are used for speech signal decomposition, those transformations assume that it is piecewise stationary and decomposition is done by fitting some predefined bases without satisfying its non-stationary nature. In this work a new technique called Empirical Mode Decomposition (EMD) is used. This technique is first introduced by N E Huang et al this technique is used to decompose any non-stationary and nonlinear signal into oscillating components called IMFs [7, 8]. The main advantage of using EMD is that it is an automatic decomposition and fully data adaptive.

In this paper sixth IMF of the speech signal is considered and the ZCR of this IMF is as shown in Figure 2. When ZCR is greater than a threshold, the selected speech segment is labeled as unvoiced and voiced otherwise.

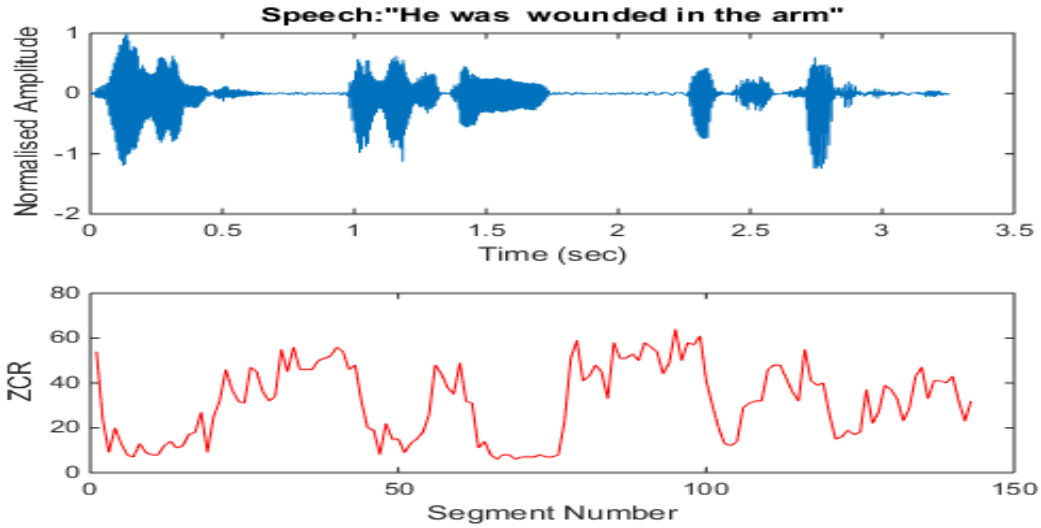

Figure 2. Speech sample 1 and its Zero crossing rate

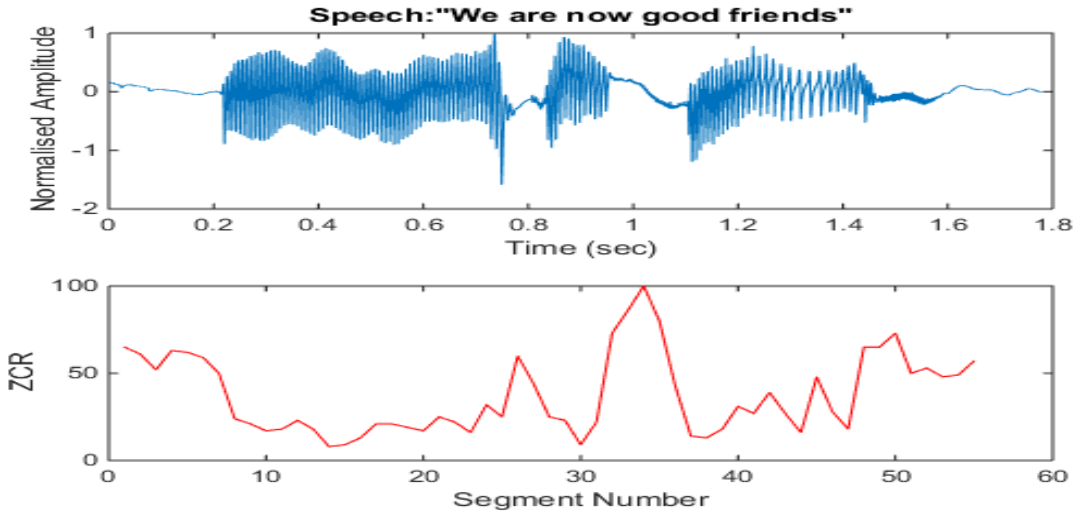

Figure 2b. Speech sample 2 and its Zero crossing rate

V Sub-band Energy (SBE)

Energy of a speech is another parameter for classifying the voiced non-voiced parts of speech. The voiced part of the speech has high energy because of its periodicity and the unvoiced part of speech has low energy so this parameter can be used to differentiate voiced and unvoiced segments in speech [12].

In this paper, wavelet sub-band energy has been proposed as one of the parameter for classification of speech signal into voiced non-voiced segments. The Discrete Wavelet Transform (DWT) is applied to extract the wavelet coefficients of the speech signal.
From the obtained wavelet coefficients sub-band energy SBE is calculated as given in equation.

$$SBE = \sum_{i=1}^{K} |C_i|^2 \quad (3)$$

Where C represents wavelet coefficients, i represent sub-bands and K is the total number of coefficients in the sub-band. This sub-band energy is used as one parameter to classify the speech segments into voiced and non-voiced components as shown in Figure 3.

Figure 3a. Speech sample and its Subband Energy

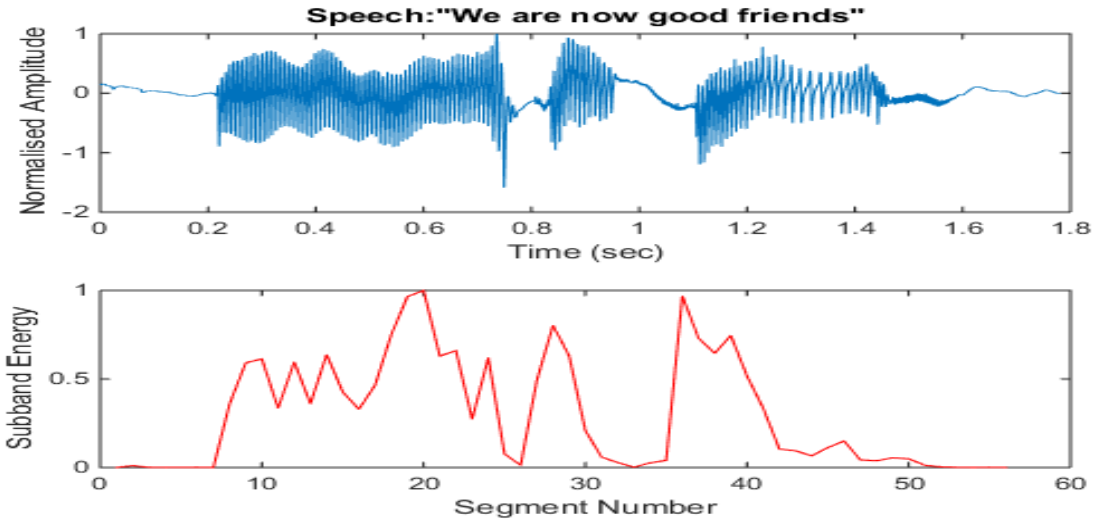

Figure 3b. Speech sample 2 and its Subband Energy

VI Short time AMDF(ST-AMDF)

The Average Magnitude Difference Function (AMDF) [17] is a variation of Autocorrelation Function (ACF). In ACF input signal is correlated at various delays and it involves multiplications and summations. But, AMDF is implemented with subtraction, addition, and

absolute value operations, calculations require no multiplications, this is a desirable property for real-time applications.

Short-time Average Magnitude Difference Function is defined as:

$$r_w(k) = \sum_{m=0}^{N-k-1} |x_w(m+k) - x_w(m)| \quad (4)$$

Plot of AMDF for different speech samples is as shown in Figure 4

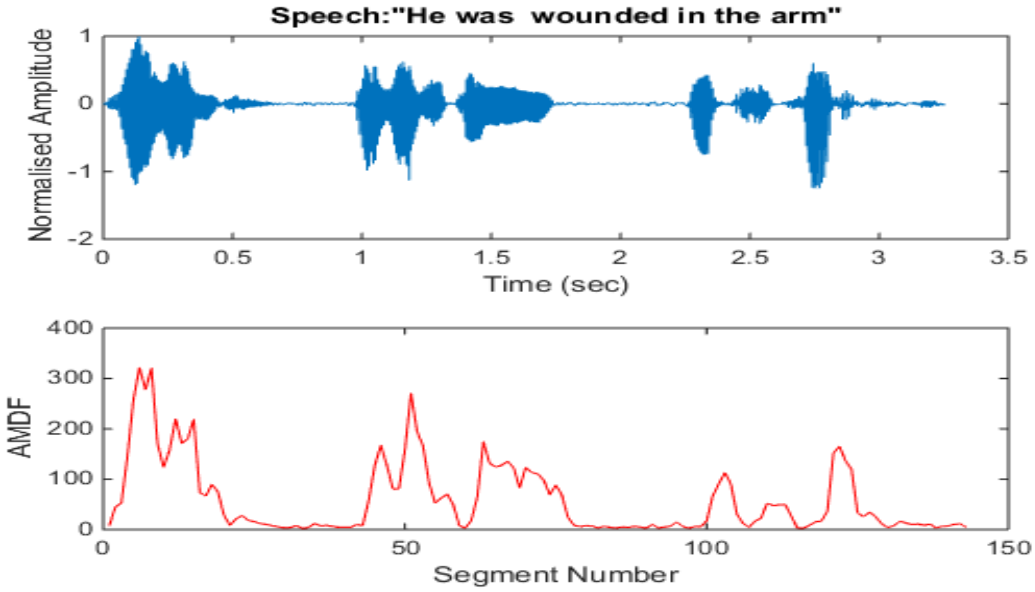

Figure 4a. Speech sample 1 and its AMDF

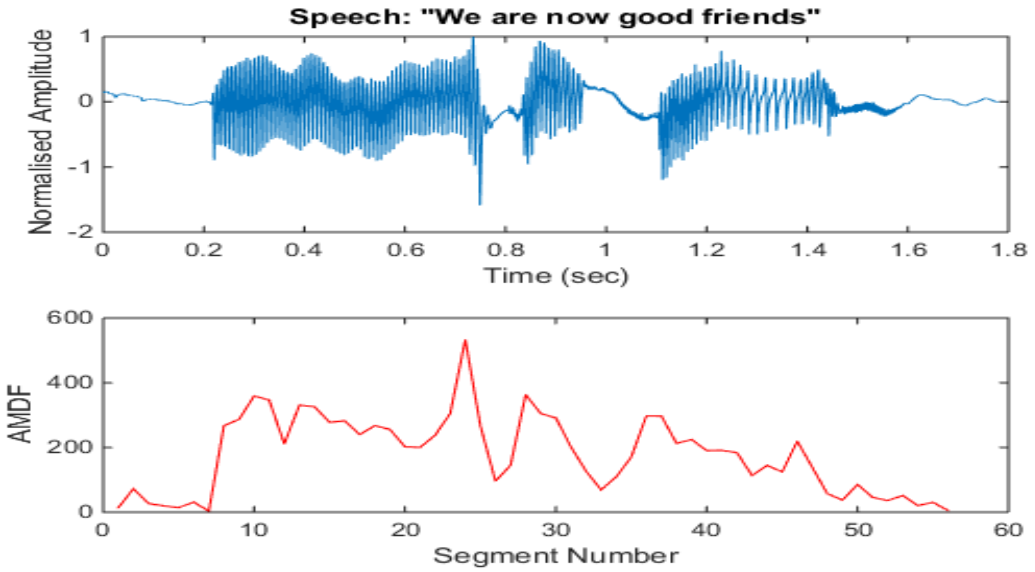

Figure 4b. Speech sample 2 and its AMDF

VII Support Vector Machine

Support Vector Machines (SVM) are set of related supervised learning methods used for classification and regression [18]. The Support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms.

They belong to a family of generalized linear classification. A special property of SVM is that they simultaneously minimize the empirical classification error and maximize the geometric

margin. So, SVMs are called maximum margin classifiers. SVMs map input vector to a higher dimensional space where a maximal separating hyperplane is constructed, then SVM finds a linear separating hyperplane with the maximal margin in this higher dimension space.

SVM has shown good performance in data classification. Its success depends on the tuning of several parameters to select the best parameter set. Then apply this parameter set to the training dataset and then get the classifier. After that, use the classifier to classify the testing dataset as shown in Figure 5.

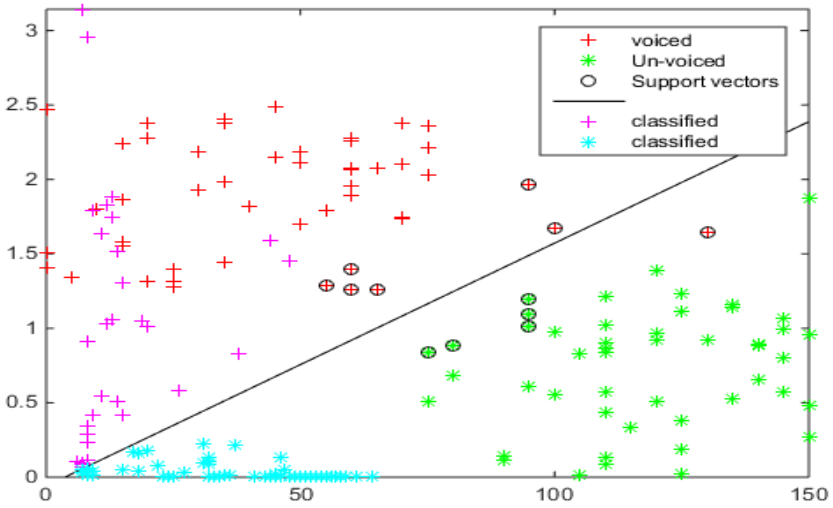

Figure.6 SVM plot to classify the speech samples

Table 1: Voiced/unvoiced decisions speech samples using different algorithms.

Speech Signal	Decision							
	IMF ZCR		Sub-band Energy		AMDF		SVM	
	Voiced	Non voiced	Voiced	Non voiced	Voiced	Non voiced	Voiced	Non voiced
"He was wounded in the arm "	46	12	47	11	52	6	50	8
"He left again and the Club caught him once more "	80	18	54	44	73	25	75	23
"We are now good friends "	53	3	39	17	39	17	41	15

Conclusion

The main purpose of this paper is to propose the use of the multiple speech parameters for classification of speech signals into voiced and non-voiced segments to make the system more noise-robust and accurate. So, in this work three features are extracted by analyzing the signal in time-frequency domain and these features are given as input to SVM model for training the classifier. The classifier is used to classify the test signal into voiced and non-voiced segments.

The comparative study of classification accuracy between using single parameter and SVM based method using many parameters is shown in Table 1

It is observed that, the proposed method attains increased reliable identification results and better detection accuracy than the existing methods. The classification accuracy

obtained by using proposed method is 98.33%.

References

1. Dix, A.J., Finlay,J., Abowd, G., Beale, R. (1998). Human-Computer Interaction, 2nd edition, Prentice Hall, Englewood Cliffs, NJ,USA.
2. Clark, J.E. and Yallop, C., (1995), *An Introduction to Phonetics and Phonology*, Blackwell, Oxford.
3. C. Hoelper, A. Frankort, C. Erdmann. "Voiced/unvoiced/silence classification for offline speech coding", International student conference on electrical engineering, Prague, 2003
4. H. Deng, D. O'Shaughnessy, "Voiced-Unvoiced-Silence Speech Sound Classification Based on Unsupervised Learning", In proceedings of IEEE International Conference on Multimedia and Expo, pp. 176-179, July 2007.
5. N. Dhananjaya, B. Yegnanarayana, "Voiced/Nonvoiced Detection Based on Robustness of Voiced Epochs," IEEE Signal Processing Letters, vol.17, no. 3, pp. 273-276, March 2010
6. Thomas F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, MIT Lincoln Laboratory, Lexington, Massachusetts, Prentice Hall, ISBN-13:9780132429429.
7. K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," Electronics Letters, vol. 36, no. 2, pp. 180–181, 2000.
8. N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," Consumer Electronics, IEEE Transactions on, vol. 57, no. 1, p. 196, 2011.
9. T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in Conference of the International Speech Communication Association, pp. 369–372, 2005.
10. S. Basu, "A linked-HMM model for robust voicing and speech detection," in Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE In-ternational Conference on, vol. 1, pp. I–816–I–819, 2003.
11. P. Mermelstein, "Automatic Segmentation of Speech into Syllabic Units", Journal of Acoustical Society of America, Vol. 58, No. 5, October 1975, pp. 880-883.
12. Bachu, R.G., Kopparthi, S., Adapa, B., Barkana, B. D., Separation of Voiced and Unvoiced using Zero-Crossing Rate and Energy of the Speech Signal. American Society for Engineering Education (ASEE) Zone Conference Proceedings, 2008.
13. Sarvajith M, BrijeshKumar Shah, Dushyanth. N .D, S. Jana "A Novel Approach for Bearing Fault Detection and Classification using Acoustic Emission Technique", UACEE International Journal of Advances in Electronics Engineering Volume 2: Issue 2 ISSN 2278 – 215.
14. N. Emanet, "ECG Beat Clas sification by Using Discrete Wavelet Transform and Random Forest Algorithm", Conference Publications, 2- 4 Sept. 2009, pp 1-4, Istanbul, Turkey
15. Pachori, R.B., Bajaj, V.: Analysis of normal and epileptic seizure EEG signals using empirical mode decomposition, Computer Methods Progr. Biomed., in press, (2011). doi:10.1016/j.cmpb.2011.03.009
16. Huang, N.E., et. al. The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lon. A 454, 903-995 (1998)
17. Dong J., Wei X., Zhang Q., and Zhao Y., "Pitch Detection Using Circular Average Magnitude Difference Function Based on Wavelet Transform," International Journal of Innovative Computing, Information and Control, vol. 5, no. 9, pp. 2717-2724, 2009.
18. Cristianini, N., and Shawe-Taylor, J. "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", First Edition (Cambridge: Cambridge University Press) (2000)