Middle-East Journal of Scientific Research 19 (2): 184-190, 2014
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2014.19.2.11469

Voiced/ Unvoiced Speech Discrimination Using Symbolic Dynamics

Chandrakar Kamath

Department of Electronics and Communication,
Manipal Institute of Technology, Manipal-576104, India

Abstract: The aim of this study is to evaluate how far the nonlinear symbolic dynamics approach helps to characterize the nonlinear properties of speech and thereby discriminate between voiced and unvoiced speech segments. The symbolic dynamics calculations were performed on voiced speech, unvoiced speech and silence data. Differences were found in histogram properties and complexity measures of symbol sequences among the three groups. The results of the analysis suggest that the nonlinear symbolic dynamics approach is helpful in classification of speech segments.

Key words: Speech · Symbolic dynamics · Voiced and unvoiced

INTRODUCTION

The need for deciding whether a given segment of speech signal is voiced, unvoiced or silence provides an important basis for speech processing applications, such as speech enhancement, speech recognition, speech coding and speech synthesis. In the recent years considerable efforts has gone in addressing this problem and a variety of approaches have been proposed to extract features for making this decision [1-10]. In [1, 2], statistical parametric methods have been used while in [3-5], non-parametric methods have been used. Acoustical features and pattern recognition techniques are used in [6] to discriminate between voiced and unvoiced speech segments. [7-9] employ zero-crossing rate and short-time energy to separate voiced/unvoiced speech. However, in [9] additional features like Teager energy and entropy are added for a better decision criteria. A new algorithm for voiced/unvoiced speech discrimination in noise is developed in [10] using Gabor atomic decomposition.

Physiological data more often show complex structures which can not be quantified using linear methods. The classical nonlinear methods suffer from the disadvantage of dimensionality. Further, there are not enough samples in the time series to arrive at a reasonable estimate of the nonlinear measures. From this point of view it is advisable to resort to methods which can quantify system dynamics even for short time series, like the symbolic dynamics. The prime advantage of symbolic dynamics is the following: If two time series have different standard deviations then the time-domain (TD) parameters can distinguish between the two time series. If two time series have different power spectra then the frequency-domain (FD) parameters can distinguish between the two time series. If two time series have the same standard deviation and power spectra then the parameters, either in TD or FD can not readily distinguish between the two time series. However, the symbolic dynamics can clearly distinguish between the two (for example between the original series and its surrogate series). Other advantages of this analysis include increase in efficiency of numerical computations compared to what it would be for original data and lower sensitivity to measurement noise. Symbolic time series analysis has found application for the past few decades in the field of complexity analysis, including cardiology (particularly, heart rate variability) [11-14], encephalography [15], combustion [16] and multiphase flow [17], astrophysics, geomagnetism, geophysics, classical mechanics, medicine and biology, plasma physics, robotics, communication and linguistics [18]. In this work we employ symbolic dynamics to decide whether a given segment of speech is voiced, unvoiced or silence. A unique feature of this new approach is using nonlinear complexity (symbolic dynamic) analysis as a more direct and perhaps, more sensitive, measure of the degree of "chaos" in speech segments to discriminate between voiced and unvoiced speech.

Corresponding Author: Chandrakar Kamath, Department of Electronics and Communication,
Manipal Institute of Technology, Manipal-576104, India.

MATERIALS AND METHODS

Analyzed Data: The performance of the proposed voiced/unvoiced classification algorithm using symbolic dynamics is evaluated using 50 sentences (sampled at 8,000 Hz) uttered by male/ female speakers. The data is manually segmented into voiced, unvoiced and silence segments with an identical frame size. In this work each frame has 80 samples.

Symbolic Dynamics

Static and Dynamic Transformations: Symbolic dynamics, as an approach to investigate complex systems, facilitates the analysis of dynamic aspects of the signal of interest. The concept of symbolic dynamics is based on a coarse-graining of the dynamics [19]. That is the range of original observations or the range of some transform of the original observations such as the first difference between the consecutive values, is partitioned into a finite number of regions and each region is associated with a specific symbolic value so that each observation or the difference between successive values is uniquely mapped to a particular symbol depending on the region into which it falls. The former mapping is called static transformation and the latter dynamic transformation. Thus the original observations are transformed into a series of same length but the elements are only a few different symbols (letters from the same alphabet), the transformation being termed symbolization. A general rule of thumb is the partitions must be such that the individual occurrence of each symbol is equiprobable with all other symbols or the measurement range covered by each region is equal. This is done to bring out ready differences between random and nonrandom symbol sequences. The transformations into symbols have to be chosen context dependent. For this reason, we use complexity measures on the basis of such context-dependent transformations, which have a close connection to physiological phenomena and are relatively easy to interpret. This way the study of dynamics simplifies to the description of symbol sequences. Some detailed information is lost in the process but the coarse and robust properties of the dynamic behavior is preserved and can be analyzed [19]. After symbolization the next step in the identification of temporal patterns is the construction of symbol sequences of specific length L, termed words, from the symbol series by gathering groups of symbols in the temporal order. L is called the word length. This sequencing process involves definition of a template of finite length L that can be moved along the symbol series one symbol at a time, each step revealing a new sequence/ word. If each possible new sequence is identified by a unique identifier the resulting series will be a new time series, termed word-sequence series. The next step is to evaluate the relative frequency of occurrence of all possible words. A simple way to keep track word-sequence frequencies is to assign a unique value, called symbolic code, to each word by computing the corresponding base-10 value for each base-n word, where, n is the number of partitions. The next step is to plot symbol-sequence frequencies as a function of symbolic code, the plot being termed symbol-sequence histogram. Because of the above rule of thumb for partitioning, for a truly random data the relative frequency of all possible symbolic codes will be equal. This implies that any significant deviation from this equiprobable feature is an indication of deterministic characteristic of the given data, the more the deviation the more is the data deterministic and time correlated.

For example, given the series $x_1, x_2, x_3, \ldots x_N$. In the static transformation [12], assuming uniform quantization, the full range of the series is spread over ξ symbols with a resolution of $(x_{max}-x_{min})/\xi$, where, x_{max} and x_{min} are respectively the maximum and minimum of the series, x. After quantization the series x becomes a new series $x_\xi = \{x_\xi(i), i=1,2, \ldots N\}$ of integer values ranging from 0 to ξ-1. Then this series is transformed into a new series, $x_{\xi,L} = \{x_{\xi,L}(i), i=1,2, \ldots N\}$, depending on a sequence of patterns of L delayed samples, where, $x_{\xi,L}(i) = \{ x_\xi(i), x_\xi(i-1), x_\xi(i-2), \ldots x_\xi(i-L+1)\}$. The number of possible $x_{\xi,L}(i)$ is ξ^L.

Symbolic Dynamics and Voiced/ Unvoiced Speech: In this study, we use yet another symbolic dynamics approach, which is basically a dynamic transformation, where a sliding window of length corresponding to five values of the series x is shifted (with τ=1), one element at a time, over the entire series as in the Eq. (1) below and the symbol s_k is computed [20]. Within each window the number of consecutive x_i differences ($|x_i-x_{i+1}|$) that fall below a scaled (in this work scaling factor=1) standard deviation of the five x_i's in the current window, is counted and coded as a symbol s_k. This process results in a symbol string with a range of five possible symbols {0, 1, 2, 3, 4}. We adopt this new approach to symbolic dynamics because the differenced symbolization scheme is relatively insensitive to extreme noise spikes in the data. By comparing different kinds of such transformations, we empirically found that the use of five symbols is appropriate for our purpose.

$$S_k = \sum_{i=1}^{i=4} \begin{cases} 0 : \left| X^k{}_i - X^k{}_{i+1} \right| >= a * sd(k) \\ 1 : \left| X^k{}_i - X^k{}_{i+1} \right| <= a * sd(k) \end{cases} \quad \text{k=1:L} \tag{1}$$

L represents the total number of shifts/ windows required to cover the entire original time series. Eq. (1) leads to a symbol from the alphabet {0, 1, 2, 3, 4}, each indicating a unique pattern. For example, a '0' implies no pair of adjacent elements in the current window with the magnitude of difference between them less than the scaled standard deviation. A '1' implies one pair of adjacent elements in the current window with the magnitude of difference between them less than the scaled standard deviation. Likewise, a '2' implies existence of two pairs of adjacent elements in the current window with the magnitude of difference between them less than the scaled standard deviation and so forth.

There are several quantities that properly characterize such symbol strings. In this work we investigate the frequency distribution (relative frequencies) of each of the patterns/ symbols from the alphabet {0, 1, 2, 3, 4}, plot the symbol histogram and perform pattern classification. We also investigate the frequency distribution (relative frequencies) of length 2 words, i.e. substrings which consist of two adjacent symbols from the alphabet {0, 1, 2, 3, 4} leading to a maximum of 25 different words/ bins. This is a compromise between retaining important dynamical information and of having a robust statistics to estimate probability distribution. We also plot word sequence (length 2) histogram to evaluate some parameters explained in the next section.

Measures of Complexity: The first measure of complexity is the Shannon entropy defined below [13]. A larger value implies higher complexity and a smaller value implies a lower complexity. From the probabilities $p(s^k)$ of words of length k we evaluate k^{th} order Shannon entropy as given by

$$H_k = - \sum p(s^k) \log(p(s^k)) \text{ for } s^k \text{ and } p(s^k) > 0 \tag{2}$$

The second measure of complexity is simply counts of number of 'forbidden words', those words which never or almost never occur [13]. We counted the number of words which never occurred or rarely occurred, in our case with probability less than or equal to 0.0015. It is important to observe that for a given dynamical system all sequences are not realizable. A large number of these words mean a reduced dynamic behavior of the time series and a smaller number means a higher complexity of the series.

The third measure of complexity is the parameter 'wsdvar' [19], which measures the variability of the time series based on the specified word sequence. Suppose that the word sequence is $\{w_1, w_2, w_3, \ldots\}$ with length 3 words. To arrive at 'wsdvar' the following sequence $\{s'_1, s'_2, s'_3, \ldots\}$ is computed first.

$$s'_i(w_i) = \begin{cases} 3 \text{ if } n_{ab}(wi) = 3 \text{ and } s_{ab}(wi) = a' \\ 2 \text{ if } n_{ab}(wi) = 2 \text{ and } s_{ab}(wi) = a' \\ 1 \text{ if } n_{ab}(wi) = 1 \text{ and } s_{ab}(wi) = a' \\ 0 \text{ if } n_{ab}(wi) = 0 \qquad\qquad i=1, 2, 3,.... \\ -1 \text{ if } n_{ab}(wi) = 1 \text{ and } s_{ab}(wi) = b' \\ -2 \text{ if } n_{ab}(wi) = 2 \text{ and } s_{ab}(wi) = b' \\ -3 \text{ if } n_{ab}(wi) = 3 \text{ and } s_{ab}(wi) = b' \end{cases} \tag{3}$$

where $n_{ab}(w_i)$ represents the total number of symbols 'a' or 'b' in the word w_i and $s_{ab}(w_i)$ represents that symbol 'a' or 'b' which occurs first in the word w_i. The parameter 'wsdvar' is defined as the standard deviation of the sequence $s'_i(w_i)$. For word sequence with length 2 words, the top and the bottom rows corresponding to $s'_i(w_i) = 3$ and $s'_i(w_i) = -3$ do not occur. The symbols 'a' and 'b' are the dominant symbols and depend on the class of speech segment, voiced or unvoiced. A higher percentage of words containing these symbols is a good measure of the respective class and is reflected in 'wsdvar'.

The next pair of measures of complexity is 'plvar10' and 'phvar10', which respectively represent low and high variability from successive symbols of another alphabet, say, B = {0,1}, comprising only symbols 0 and 1. The symbol 0 represents the case when successive difference between elements in the series does not exceed a specified limit, limit and the symbol 1 represents the case when successive difference between elements in the series is either equal to or does exceed a specified limit, limit, as given below [14].

$$S_k = \begin{cases} 1 : |x_k - x_{k+1}| >= \text{limit} \\ 0 : |x_k - x_{k+1}| < \text{limit} \end{cases} \tag{4}$$

Words of length 6, consisting of only 0s or only 1s are counted [19]. The former represents the probability of occurrence, 'plvar10', of the word type '000000'and the latter represents the probability of occurrence, 'phvar10', of the word type '111111'.

RESULTS AND DISCUSSION

Characterizing and Comparing Symbol Histograms: The speech data is manually segmented into voiced, unvoiced and silence segments with a varying frame size. Fig. 1 shows the voiced, unvoiced and silence segments of a speech signal. Symbolic dynamics is applied to each of these different segments to decide whether a particular segment was voiced/ unvoiced or silence. Eq. (1) is then applied on each segment to arrive at a symbol string with a range of five possible symbols {0, 1, 2, 3, 4}. Relative frequencies of these symbols/ patterns are computed over the entire speech segment and the symbol histogram is plotted for each speech segment. Fig. 2 shows these histograms for voiced, unvoiced and silence segments of speech. The distribution of patterns for the three cases is found to be distinctly different. Comparison between Fig. 2(a) and Fig. 2(b) reveals that

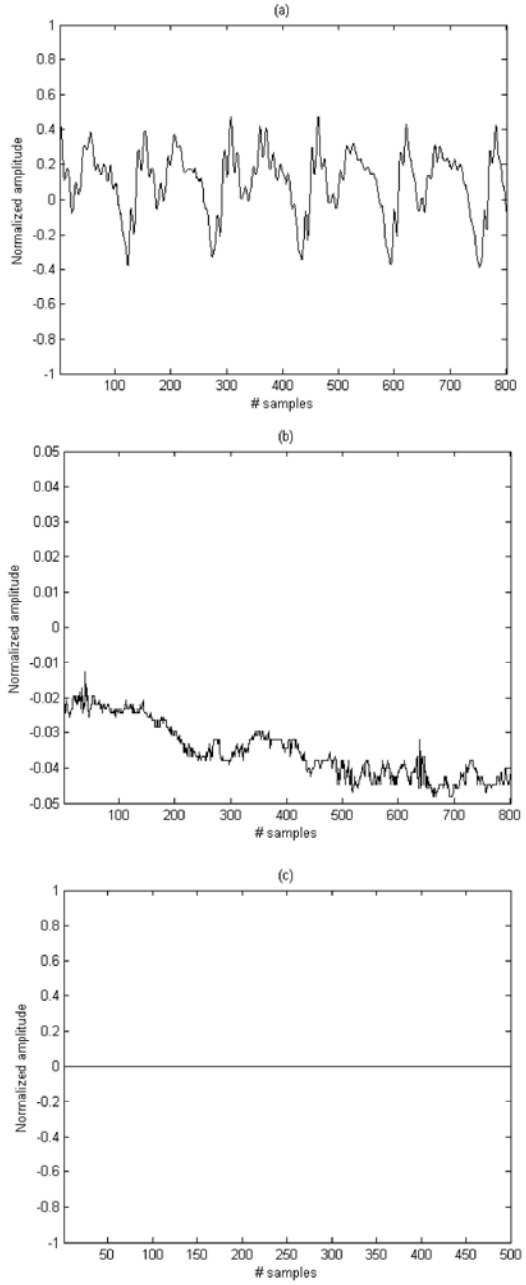

Fig. 1: Segments of speech
(a) Voiced speech (b) Unvoiced speech (c) Silence

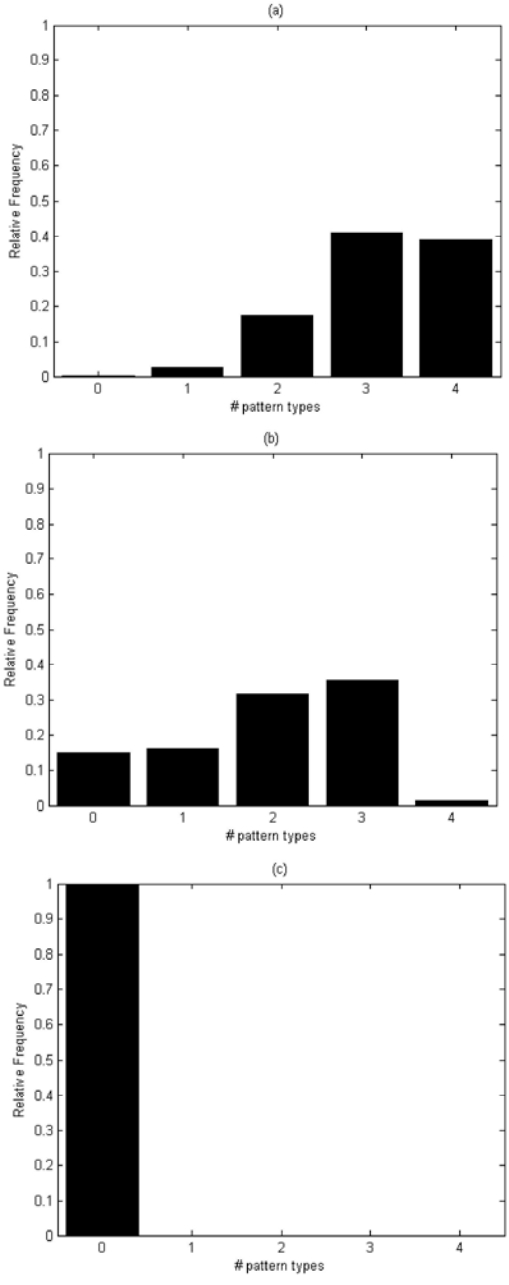

Fig. 2: Symbol histogram
(a) Voiced speech (b) Unvoiced speech (c) Silence

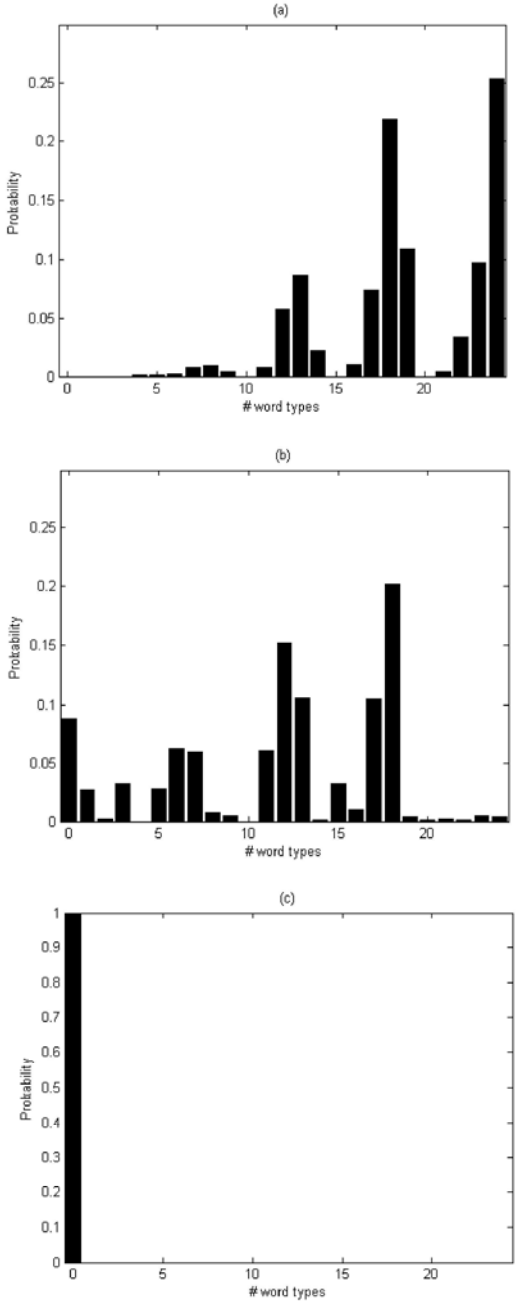

Fig. 3: Symbolic-sequence histogram
(a) Voiced speech (b) Unvoiced speech (c) Silence

for voiced speech the patterns represented by symbols 3 and 4 considerably dominate the respective symbols of unvoiced case while for unvoiced speech the patterns represented by symbols 2 and 3 comparatively dominate the respective symbols of voiced case. For perfect silence the only pattern represented by symbol 0 dominates others being 0 as implied by Fig. 2(c).

Characterizing and Comparing Symbol-sequence Histograms: From the same symbol strings, words of length 2 are built with an overlapping of one symbol. We assign a sequence code for each of the words by using equivalent base-10 value for each of the base-n word, where, n is the number of partitions. The relative frequencies of length 2 words are then computed and symbolic sequence histogram is plotted for each of the speech segments. Fig. 3 compares these word histograms for voiced, unvoiced and silence segments of speech. The distribution of patterns for the three cases reveals distinct difference. Comparison between Fig. 3(a) and Fig. 3(b) shows that for voiced speech those words containing symbols/ patterns 3 and 4 dominate compared to those words containing 0, 1 and 2. For example. the words 24, 18,19 and 23 (i.e., 44, 33, 34 and 43 respectively, in base-10) have much higher relative frequencies than those words not made of symbols 3 and 4. But for unvoiced speech the patterns represented by symbols 2 and 3 dominate compared to those words with 1 and 4. For example. the words 18, 12, 13 and 17 (i.e., 33, 22, 23 and 32 respectively, in base-10) have much higher relative frequencies than those words not made of symbols 2 and 3. Also it is found that in both the voiced and unvoiced cases words with symmetric behavior (i.e., 44, 33, 22, … in base-10) exhibit higher frequencies. For example. the words 24 and 18 in voiced case and 18 and 12 in unvoiced case. Further, it is observed that in both the cases words with diagonal behavior (i.e., [34, 43], [23, 32], in base-10) have almost identical probabilities. Fig. 3(c) shows that for perfect silence only the word 00 is prominent and all the other words are zero.

Shannon Entropy: For the voiced segment shown Shannon entropy is 0.7013 and for unvoiced segment shown it is 0.7655 implying that voiced speech more informative than unvoiced speech. Obviously, for silence it is zero.

Forbidden Words: Forbidden words are counted for each case and it is found that for voiced speech 00, 02, 03, 04, 09, 10, 20 and 21 are the usual forbidden words, for unvoiced speech 02, 04, 10, 14, 20 and 21 are the usual forbidden words and for silence all the words except 00 are the forbidden words. This means silence is the most complex, while unvoiced speech is more complex than voiced speech.

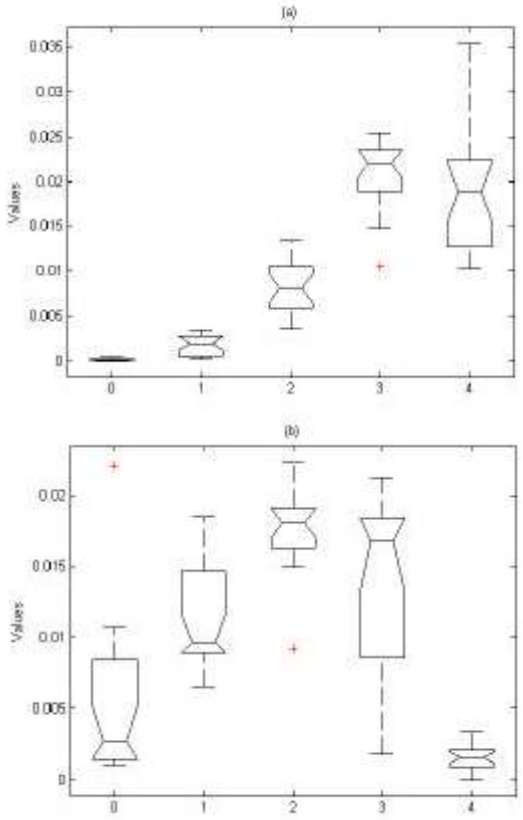

Fig. 4: Box and whisker plot for the variations of symbol patterns
(a) Voiced speech (b) Unvoiced speech

'Wsdvar': Next for length 2 words a new string $s_i'(w_i)$ is generated as per Eq. (3) and we compute 'wsdvar' for two cases: (1) with a = 3 and b = 4, dominant patterns for voiced speech and (2) with a = 2 and b = 3, dominant patterns for unvoiced speech. For voiced speech this parameter evaluates to 1.4482 for case (1) and 0.6056 for case (2). For unvoiced speech it evaluates to 0.8350 for case (1) and 0.6765 for case (2). Thus depending on dominant patterns it is possible to find whether a given speech segment is voiced or unvoiced.

'Phvar10' and 'Plvar10': With limit = 0.10 (on a normalized scale) we generate the symbol strings using definition of Eq. (4) for voiced, unvoiced and silence segments of speech. Words of length 6, consisting of only 0s or only 1s, are counted and 'phvar10' and 'plvar10' are computed for each case. For voiced speech these values respectively evaluate to 328 and 26, for unvoiced speech these values are 0 and 794, respectively and for silence they are 0 and 493, respectively. It is found that 'phvar10' is high for voiced speech and low for unvoiced speech and silence. On the other hand, 'plvar10' is low for voiced speech and high for unvoiced speech and silence.

Table 1: Statistics of the symbol patterns for voiced and unvoiced speech segments. Values are expressed as median (first quartile-third quartile).

Symbol pattern	Voiced	Unvoiced
0	9.411e-005	0.00273
	(6.007e-006 –0.000251)	(0.00141 – 0.00844)
1	0.00197	0.00960
	(0.00056 4 –0.00276)	(0.00887 – 0.01462)
2	0.00812	0.01804
	(0.00592 – 0.01045)	(0.01628 – 0.01914)
3	0.02199	0.01675
	(0.01888 – 0.02353)	(0.00862 – 0.01842)
4	0.01885	0.00156
	(0.01277 – 0.0224)	(0.00094 – 0.00207)

Statistical Analysis: Finally we also perform a visual hypothesis test for medians of symbol patterns using Box-whiskers plots. Fig. 4(a) and 4(b) show the distribution of medians of the five symbol patterns from 50 sentences for the voiced and unvoiced cases respectively, using Box-whiskers plot. None of the respective boxes (inter-quartile range) of the symbol patterns for the voiced and unvoiced cases overlap. None of the whiskers also overlap except for symbol pattern 3. In this case the lower whisker (lower quartile) of the symbol pattern 3 for the voiced case overlaps with the corresponding upper whisker (upper quartile) for the unvoiced case. The values are expressed as median (first quartile-third quartile) for the various symbol patterns in the voiced and unvoiced cases as shown in Table 1. This implies that the median values for the symbol patterns are all significantly different for both the case studies and can be readily used to discriminate between voiced and unvoiced speech.

CONCLUSION

We present a new approach to separating voiced, unvoiced, or silence part of speech using symbolic dynamic analysis. The frequency distribution in symbol histogram and symbol-sequence histogram, both reveal significant differences among the three classes. Almost all parameters from symbolic dynamics facilitate considerably the separation among the different classes of speech, namely, voiced speech, unvoiced speech and silence.

The presented results of this study show the effectiveness of symbolic dynamics in speech analysis. Preliminary results show that even in the presence of noise symbolic analysis works satisfactorily. In our future study, we plan to improve our results for voiced/unvoiced discrimination in noise.

REFERENCES

1. Atal, B. and L. Rabiner, 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Trans. on ASSP ASSP, 24: 201-12.
2. Ahmadi, S. and A.S. Spanias, 1999. Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. IEEE Trans. Speech Audio Processing, 7(3): 333-8.
3. Qi, Y. and B.R. Hunt, 1993. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. IEEE Trans. Speech Audio Processing, 1(2): 250-5.
4. Siegel, L., 1979. A procedure for using pattern classification techniques to obtain a voiced/ unvoiced classifier. IEEE Trans. on ASSP, 27: 83-8.
5. Burrows, T.L., 1996. Speech Processing with Linear and Neural Network Models, Ph.D. thesis, Cambridge University Engineering Department, UK.
6. Jashmin Shah, K., N. Ananth Iyer, Y. Brett Smolenski and E. Robert Yantorno, 2004. Robust voiced unvoiced classification using novel features and Gaussian Mixture model. Proc. of ICASSP04.
7. Bachu, R.G., S.B. Kopparthi and B.D. Barkana, 2008. Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal. American Society for Engineering Education ASEE Zone Conference Proceedings.
8. Greenwood, M. and A. Kinghorn, 1998. Suving Automatic silence/ unvoiced/ voiced classi?cation of speech. Department of Computer Science, The University of Sheffield.
9. Alexandru Caruntu, Alina Nica and Gavril Toderean, 2006. Robust features for speech classification. Fifth International Symposium on CSNDSP, pp: 297-300.
10. Lobo, A.P. and P.C. Loizou, 2003. Voiced and unvoiced speech discrimination in noise using Gabor atomic decomposition. ICASSP, 1: I-820-3.
11. Kurths, J.A. Voss, P. Saparin, A. Witt, H.J. Kleiner and N. Wessel, 1995. Quantitative analysis of heart rate variability. Chaos, 5: 88-94.
12. Porta, A.G. D'Addio, G.D. Pinna, R. Maestri, T. Gnecchi-Ruscone, R. Furlan, N. Montano, S. Guzzetti and A. Malliani, 2005. Symbolic analysis of 24h Holter heart period variability series: Comparison between normal and heart failure patients. Computers in Cardiology, 32: 575-8.
13. Eleonora Tobaldini, Alberto Porta, Wei Shun-Guang, Zhi-Hua Zhang, Joseph Francis, Karina Rabello Casali, M. Robert Weiss and B. Robert Felder, 2009. Nicola Montano. Symbolic analysis detects alterations of cardiac autonomic modulation in congestive heart failure rats. Auton Neurosci, 150(1-2): 21-6.
14. Wessel, N.U. Schwarz, P.I. Saparin and J. Kurths, 2002. Symbolic dynamics for medical data analysis. In: Attractors, Signals and Synergetics, pp: 45-61.
15. Xu, J.H., Z.R. Liu and R. Liu, 1994. The measures of sequence complexity for EEG studies. Chaos, 4(11): 2111-9.
16. Daw, C.S., 1998. Observing and modeling nonlinear dynamics in an internal combustion engine. Phys. Rev. Lett., 57(3): 2811-9.
17. Finney, C.E.A., K. Nguyen, C.S. Daw and J.S. Halow, 1998. Symbol-sequence statistics for monitoring fluidization.Proceedings of the ASME Heat Transfer Division, pp: 405-11.
18. Daw, C.S., C.E.A. Finney and E.R. Tracy, 2003. A review of symbolic analysis of experimental data. Review of Scientific Instruments, 74(2): 915-30.
19. Voss, A., J. Kurths, H.J. Kleiner, A. Witt, N. Wessel, P. Saparin, K.J. Osterziel, R. Schurath and R. Dietz, 1996. The application of methods of nonlinear dynamics for the improved and predictive recognition of patients threatened by sudden cardiac death. Cardiovasc. Res., 31: 419-33.
20. Voss, A., R. Schroeder, M. Vallverdu, I. Cygankiewicz, R. Vazquez, A. Bayes de Luna and P. Caminal, 2008. Linear and Nonlinear Heart Rate Variability Risk Stratification in Heart Failure Patients. Computers in Cardiology, 35: 557-60.