

A Highly Efficient Technique for the Detection of the Arabic Vowels

Moussa Abdallah

Department of Electronics Engineering, Princess Sumaya University
P.O.Box 1438 Al-Jubaiha 11941, Amman, Jordan

Email: moussa@psut.edu.jo

and

Mohammed Mohsen Olama

Department of Applied Science, University of Arkansas at Little Rock
Little Rock, AR, 72204, USA

Email: mmolama@ualr.edu

ABSTRACT

This paper investigates some of the tools and technologies related to speech processing in order to apply them to the Arabic speech. A new method has been introduced for the detection of Arabic vowel phonemes. Words containing vowels are pre-processed and then analyzed using nonlinear algorithms.

Keywords: Nonlinear Dynamic System, Back Propagation, SVD, And Arabic Vowels

INTRODUCTION

This paper is mainly concerned with the application of nonlinear techniques to the classification of Arabic vowel in Arabic speech. The use of nonlinear methods in conjunction with speech would appear justified; since it is generally accepted that there are a number of nonlinear mechanisms present in the speech production process.

The classification process is based on the use of a BP ANN. This nonlinear classifier can distinguish between the three Arabic vowels for either male or female speakers.

ALGORITHM DEVELOPMENT

In nonlinear processing the state-space of a dynamical system can be reconstructed from a time series of only one observed variable by using Takens theorem. This theorem states that there will be a one to one mapping between the reconstruction and the actual attractor of the underlying system (Takens, 1980). If the embedding is carried out in a space of dimension of at least m , where:

$$m \geq 2d + 1 \quad (1)$$

where d is the phase-space dimension of the original attractor, then this mapping is guaranteed. In practice, forming the reconstructed trajectory matrix is relatively simple, and involves moving a window of length m through the data to form a series of x vectors. So for the discrete time series $x_n = (x_0, x_1, x_2, \dots, x_i)$, the trajectory matrix X is:

$$X = \begin{bmatrix} x_0 & x_t & \cdots & x_{(m-1)t} \\ x_1 & x_{(1+t)} & \cdots & x_{(1+(m-1)t)} \\ x_2 & x_{(2+t)} & \cdots & x_{(2+(m-1)t)} \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} \quad (2)$$

In real systems, where noise is an issue, the reconstruction produced by time delay embedding inherits the noise of the time series, which will adversely affect the results of any subsequent

analysis. The Singular Value Decomposition (SVD) embedding technique was developed to resolve this problem (Broomhead, 1986). The principle behind it is to partition the state-space into two subspaces, one containing the signal and the other the noise. As a first step, the time delay embedding matrix X is formed with $t = 1$ sample and a window length of w . w will be referred to as the SVD window length and is chosen to be much greater than the supposed required embedding dimension. The singular value decomposition of X is defined by (Golub, 1989):

$$X = UWV^T \quad (3)$$

where W is diagonal and contains the singular values $w_0 > w_1 > w_2$ and U and V are orthogonal and contain the singular vectors associated with W . However, if the last elements of W are much smaller than their predecessors, it should be legitimate to discard the final columns, thus reducing the dimension and removing noise effects (Mees, 1987). Therefore, the reduced trajectory matrix can be written as:

$$X = X V_d \quad (4)$$

where V_d only contains the columns of V corresponding to the significant values of W .

It has been found that vowel sounds are low dimensional, and can be modeled on a 3 dimensional state-space (Kumar and Mullick, 1996). Figure 1 shows a SVD embedding for the vowel /u:/ in 3-D state-space. A SVD window length of 30 samples has been found to be adequate to ensure that the attractor is opened up at a sampling rate of 8 kHz.

Additionally, these reconstructions are pitch-synchronous, in that one revolution of the phase-space reconstruction is equivalent to one pitch period (Mann and McLaughlin, 1998). Clearly, this fact can be exploited to mark points in time separated by multiples of the pitch period using the Poincaré

dynamical systems, it replaces the flow of an n th order continuous system with an $(n - 1)$ th order discrete-time map.

EPOCH MARKING ALGORITHM

The algorithm uses the principle outlined above to mark successive epochs. When dealing with real speech signals, there are a number of practical issues to contend with. Because speech is nonstationary, the input signal must be treated on a frame-by-frame basis, within which the speech is assumed stationary. A frame length of 35 msec (280 samples at 8 kHz) is used, since voiced speech is often assumed to be approximately stationary for 30 to 45 msec. The input waveform is assumed to

be a voiced sound (vowel), and is blocked into overlapping frames with a 50% overlap between adjacent frames.

Each frame is embedded into 3D state-space using an SVD window length of 30 samples. The location of the first epoch x_{GCI} is determined by finding the minimum value during the middle 70 samples (experimental measure) in the first frame of the time-domain signal and obtaining its location. A Poincaré section is positioned normally to this point in state-space and all of the P points, $x_{(j)}$ ($1 \leq j \leq P$), those traverse are detected. These points can be found by looking for sign changes in $H(x)$:

$$H(x) = \langle h, (x_i - x_{GCI}) \rangle \quad 1 \leq i \leq N \quad (5)$$

where h is the approximated flow vector at x_{GCI} , x_k is the k th reconstruction vector in the frame of length N , and $\langle a, b \rangle$ is the scalar product between a and b . x will contain all of the epochs, since they are pitch synchronous with x_{GCI} . The flow chart shown in Figure 2 illustrates this process.

This technique appears useful. It works very well on the simple cases of stationary vowels and rising pitch vowels, accurately marking all the epochs, as shown in Figures 3 and 4. The algorithm is often able to track considerable changes in the phase-space structure, caused by changes in vowel sound and/or pitch.

THE CLASSIFIER

The BP ANN (Haykin, 1999) is used to recognize the three Arabic vowel phonemes for each female/male speaker. The BP ANN used in this research consists of two-layer neural network (the hidden and the output layer). The hidden layer consists of 15 neurons and tan-sigmoid transfer function. The output layer consists of one neuron and linear transfer function, so the output of this BP ANN is a single element that may take any value (Demuth, 1998). The input vector is the pitch contour output from the Epoch Marking algorithm, which contains 50 elements. If the number of elements in the pitch contour is greater than 50 elements, then the first 50 elements is taken to the input of the BP ANN. Otherwise, if the number of elements in the pitch contour is less than 50 elements, then it is zero padding to 50 elements, and the zero padding vector is the input of the BP ANN. All subjects are represented in the training dataset. The entire dataset consists of six subjects (three females and three males), and the subdivision into training and testing datasets is shown in Table 1 (a).

The classification of female/male speaker is based on the pitch contour level, which is around 250 Hz for female speaker and around 140 Hz for male speaker. So, a threshold level (195 Hz) is enough for female/male decision-making process (equally likely inputs).

After the training process is applied, the processing is performed on the test data. A flowchart summarizing the steps required to get a classification decision on the test data is shown in Figure 5.

DISCUSSION AND RESULTS

Initially the recording was performed in general room environment (laboratory room in the department). Six subjects were recorded, three males and three females, each subject read thirty utterances (words). The subjects were requested to extend the vowel portion of each utterance. The speech data were obtained by sampling at 8 kHz and digitized at 8-bit resolution (A/D conversion). The speech waveform is firstly segmented to remove other phonemes and silence, and obtain the required

phoneme alone, and then it is analyzed. The results of classification on the test dataset are shown in Table 1 (b).

The results shown above indicate that it was in fact not very difficult to obtain a high rate of classification after preprocessing. However, the results obtained are still very promising for the reason that only 20% of the dataset were used for training, and the classifier was exercised using the remaining 80% of the data.

CONCLUSIONS

Clearly the major drawback of the algorithm is that no method of locating the first epoch in the state-space domain was found. Thus, as a stand-alone technique, the algorithm is only capable of marking pitch synchronous points. Results on vowel phoneme show that the algorithm gives reasonable results, although further work is required to improve robustness, particularly when the state-space reconstruction structure is complex and this is the main reason of failure in this algorithm.

The algorithm is often able to track quite considerable changes in the attractor structure, caused by changes in vowel sound and/or pitch, but if an error occurs (usually caused by a

structure is very complicated), it is unable to recover leading to incorrect marking at all points forward from the error point. This is again due to the fact that we are not actually locating the epoch pulses specifically, only points which are pitch synchronous. Therefore, when an error occurs, causing a loss of synchronization, it propagates through the remainder of the signal. Further work is needed to address this problem, as well as improving the algorithm so it is more able to cope with sudden changes in attractor structure.

REFERENCES

- Broomhead, D.S. and King, G.P. 1986. Qualitative Analysis of Experimental Dynamical Systems, In: *Nonlinear Phenomena and Chaos*, pp. 113-144. Bristol: Adam Hilger.
- Demuth, H. and Beale, M. 1998. *Neural Network Toolbox for Use with MATLAB*, users guide version 3.0.
- Golub, G.H. and Van Loan, C.F. 1989. *Matrix Computations*, Second Edition. The John Hopkins University Press, Baltimore and London.
- Haykin, S. 1999. *Neural Networks, A Comprehensive Foundation*. Second Edition. Reading, MA: Prentice Hall, USA.
- Kumar, A. and Mullick, S. 1996. Nonlinear Dynamical Aspects of Speech, *Journal of the Acoustical Society of America*, vol. 100, pp. 737-793.
- Mann, I. and McLaughlin, S. 1998. A Nonlinear Algorithm for Epoch Marking in Speech Signals Using . *Proceedings of the 9th European Signal Processing Conference*, vol. 2, pp. 701-704.
- Mees, A., Rapp, P. and Lennings L. 1987. Singular Value Decomposition and Embedding Dimension, *Physical Review A*, vol. 36, pp. 340-346.
- Takens, F. 1980. Detecting Strange Attractors in Turbulence. *Proceedings of Symposium on Dynamical Systems and Turbulence*, pp. 366-381.

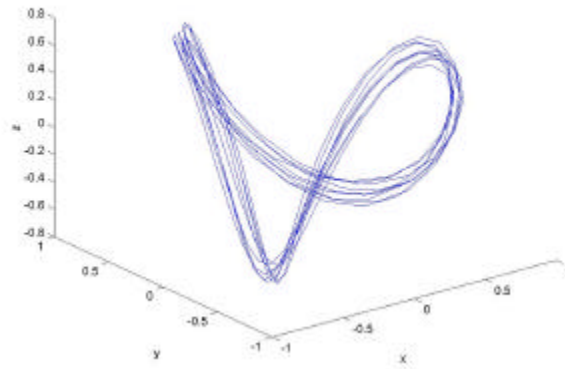


Figure 1: SVD embedding ($w = 30$) for the vowel /u:/ spoken by a female speaker.

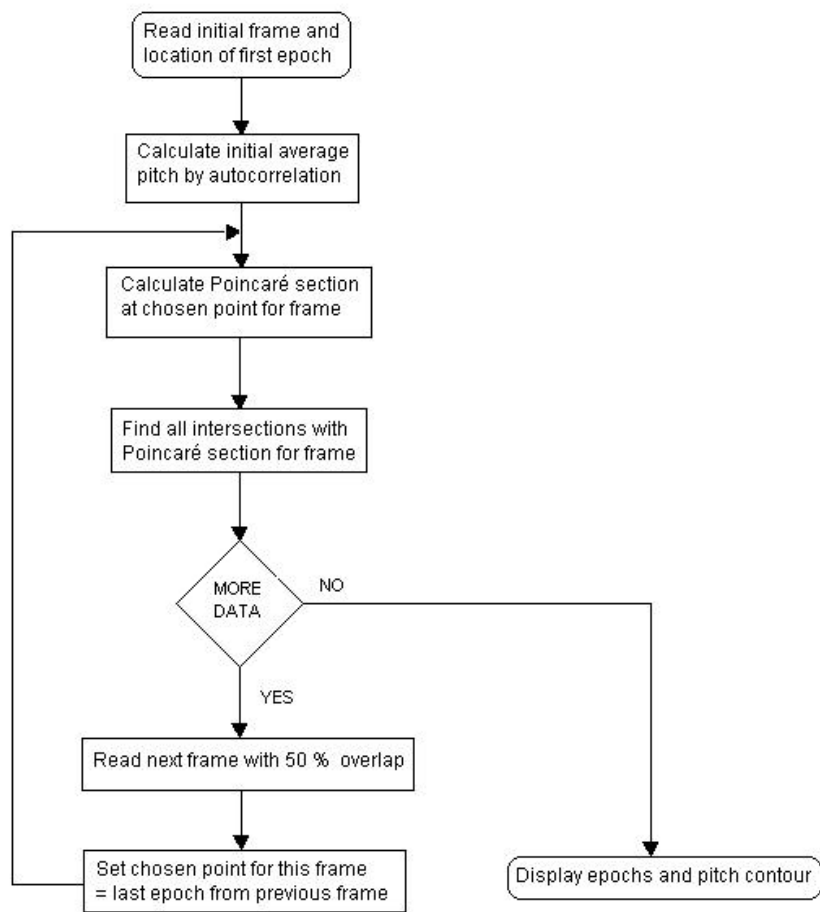


Figure 2: Schematic (flowchart) of the epoch marking algorithm

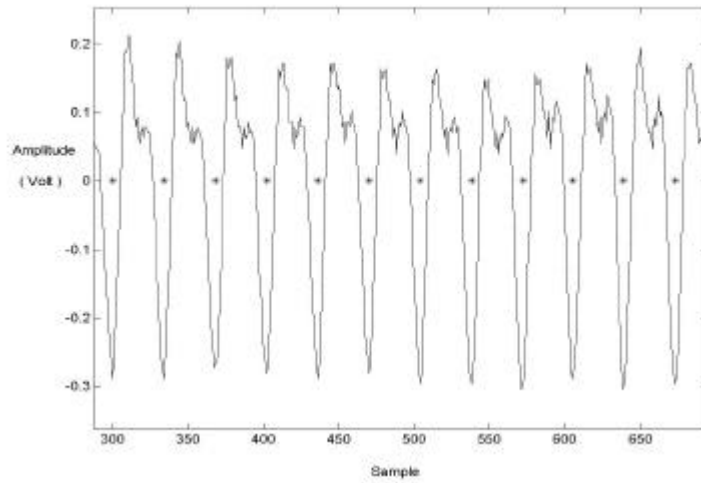


Figure 3: Results of the algorithm, the epochs shown as * on the waveform, part from the vowel /u:/ for male speaker.

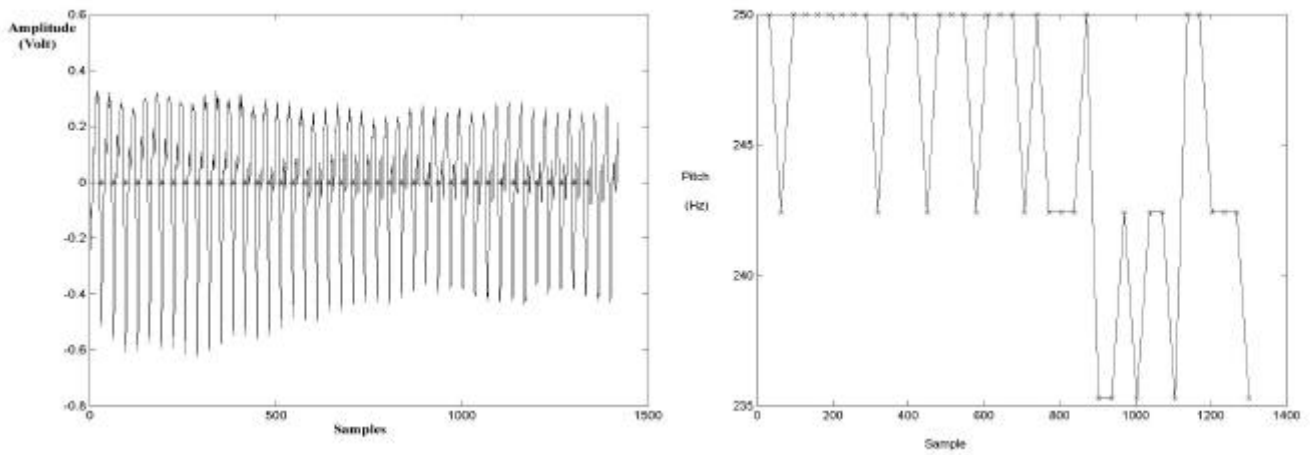


Figure 4: Results for the vowel /u:/ spoken by a female speaker, (a) the signal and the epochs as calculated b

Table 1: (a) Subdivision of data into Training and Testing Sets. (Note that over 80% of data have been reserved for testing). (b) Classification results generated from test data. (Numbers represent number of test vowels, equivalent percentage are given in parentheses).

Vowel Type	# of Training Vectors	# of Testing Vectors	Vowel Type	# of True Detection	# of False Detection
Female /a:/	5	28	Female /a:/	24	4
Female /u:/	5	21	Female /u:/	16	5
Female /i:/	5	21	Female /i:/	18	3
Male /a:/	5	28	Male /a:/	27	1
Male /u:/	5	17	Male /u:/	16	1
Male /i:/	5	25	Male /i:/	24	1
Total	30	140	Total	125 (90%)	15 (10%)

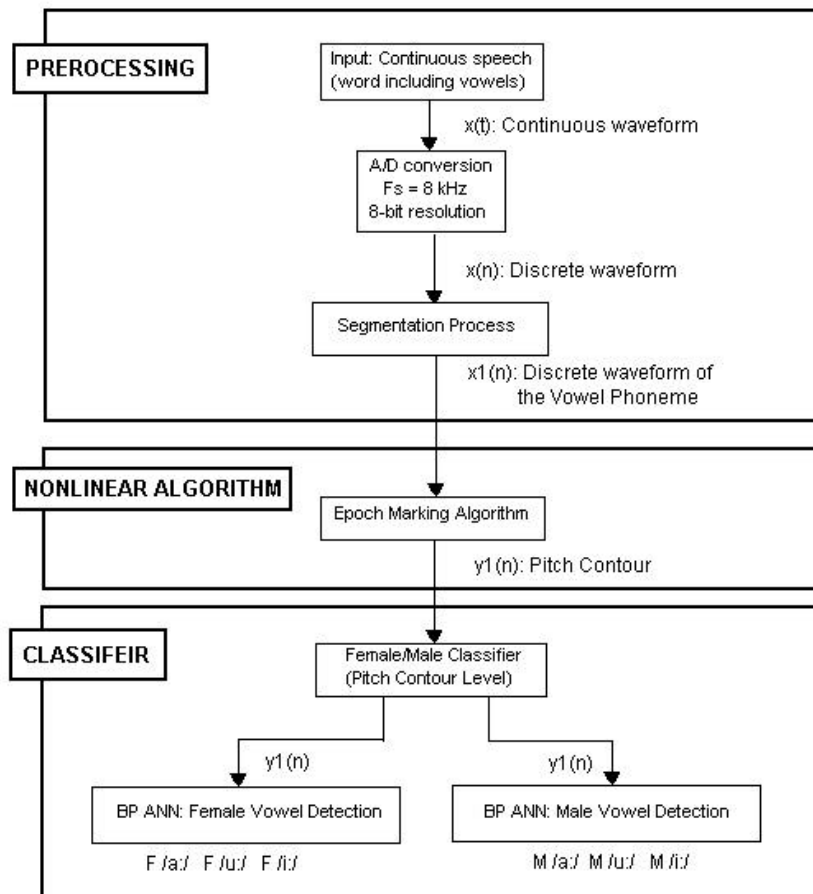


Figure 5: Components of Arabic Vowels Recognition System