Classification of Speech Signal using Phase Space Analysis – A case study in Bangla

Arup Saha, Bhaskar Gupta, Asoke Kumar Datta

Abstract—The main objective of this paper is to classify the speech signal into its basic components viz. quasi-periodic, quasi-random, and quiescent using a time-domain based state phase analysis algorithm. The uniqueness of the phase space algorithm is the conversion of a single dimension speech signal into the two-dimension array for the construction of the phase plot. These phase plots are later on used for the building of a deviation graph from where four classification parameters have been derived. As a case study of our algorithm we have selected the Bangla language for training and testing purposes. All the speech signals taken for this experiment have been recorded at the sampling frequency of 8 kHz with a 16-bit encoding system at a studio environment. The recording of the normal Bangla sentences has been done by 8 speakers across different age groups. A comparative study of the classification of the basic class of speech signal using Naïve Bayes' classification algorithm and k-NN classifier with weighted Euclidean as a distance function is also done. It has been observed that there is a significant improvement of classification results from 95.5% to 97.5% on using Bayes' classification algorithm over the k-NN classifier with weighted Euclidean as a distance function. The reduction of one parameter for classification along with the application of a maximum voting algorithm brings up the classification result further to 98.6%. The results seem to be encouraging and are expected to be replicated for other spoken languages also.

Index Terms—Phase Space Analysis, Speech Recognition, Deviation Plot, Phase Plot, Naïve Bayes Classification, Quasi Periodic, Quasi Random, Quiescent, Time Domain Analysis

1 INTRODUCTION

IN today's world, technology for man-machine interaction is gaining more and more importance in the life of people. The major reasons for such a rise in the prominence of man-machine interaction are due to the immense enhancement in computation power, digital connectivity, and efficient algorithm related to the popularization of artificial intelligence. Moreover, it is also proving to be a great tool for the accessing and dissemination of much-needed information to the functionally illiterate persons who are till now negated of the fruit of the digital revolution.

The speech processing unit is one of the important components in the design of the human-computer interface as the information content of speech is very high in comparison to other communicating media [1]. To make the human-computer interface system more effective, the speech recognition system has to be adequately efficient [2]. The last few decades have witnessed the development of various time domain, frequency domain, and hybrid algorithms for speech recognition.

In order to increase the efficiency of the recognition system, a bottom-up approach has been adopted here where the identification of the speech signal in the form of its corresponding text is done through a sequence of combination and refinement of knowledge base with speech parameters at various stages of processing [3]. It has been observed through a series of experiments that human beings recognize speech from the signal using the bottom-up approach [4]. In the bottom-up approach, the main thrust lies in the identification of phonetic units viz. quasi-periodic rather than on the phone viz. /a/ itself [3]. In order to achieve the above objective, one has to identify the following basic classes of the signal viz. quasi-periodic (QP), quasi-random (QR) and quiescent (Q) with the highest percentage of recognition so that the overall recognition rate increases. From here onwards the basic component of the signal viz. quasi-periodic, quasi-random, and quiescent will be referred to as Q3. In this regard, several algorithms have been developed using both time domain and frequency domain parameters [5]. One of the main advantages of using the time domain parameters is its simplicity of being extracted from the signal itself. On the other hand, they are liable to get corrupted in presence of the large noise [6], and hence researchers turned their attention towards the development of algorithms in the frequency domain [7]. The major disadvantage of using frequency-domain parameters is its complex process of extraction despite it being robust in presence of noise. To overcome this problem, the researchers started looking into time-frequency domain parameters i.e. parameters related to both time and frequency domain [8][9][10][11]. In this approach, the researchers are using time-frequency domain parameters in the neural networks to achieve the desired result [12].

As mentioned in the previous paragraph, one of the main usages of the Q3 component of the signal is in the application of the speech recognition system where the phonemes are being recognized through a bottom-up approach. Apart from speech recognition, the Q3 component of the speech signal is used in the speaker identification and verification system by

- *Arup Saha is currently working in speech processing lab of CDAC-Kolkata, Kolkata, West Bengal, India. E-mail: arupmtech@gmail.com*
- *Bhaskar Gupta is currently a professor in the Department of Electronics and Telecommunications at Jadavpur University, West Bengal, India. E-mail: gupta_bh@yahoo.com*
- *Asoke Kumar Datta is Ex Head of Electronics and Communication Sciences Unit and Electromechanical, Laboratory, Indian Statistical Institute Kolkata, West Bengal, India. Currently he is working as an independent researcher in the field of Speech Technology. E-mail: dattashoke@gmail.com*

comparing the segmental feature of the speech with the original speaker. Similarly, in the language recognition system, the simple frequency distribution of the Q3 component of the signal can be used to group the languages according to their family with the highest degree of recognition [13]. Moreover, the proper classification of the Q3 component of the signal can be used as an initial stepping stone for the rapid development of speech resources like annotated corpora for further research in speech technology.

In this paper we have tried to develop a time-domain process of automatic segmentation of the raw speech signal into its constituents viz. quasi-periodic, quasi-random, and quiescent (Q3). So far, we have seen that there is no such single algorithm available in the time domain by which one can get all three basic components of the signal at the same time. This very fact motivated us to develop a time-domain signal processing unit that can automatically label the signal into its constituents Q3 with the highest degree of recognition rate so that it is useful in a bottom-up approach for speech recognition. One of the main highlights of this algorithm is the conversion of a single dimension speech signal into a low dimensional array of parameters. A comparative study for classification of the basic constituent of the signal using Euclidean distance classifier and Bayes classifier have been explored. The current study has been conducted on a large set of Bangla speech database [14].

The Bangla language is one of the most widely spoken languages in the world. According to a current survey, it ranked fifth as the most spoken native language of the world [15][16] with approximately 228 million people as a first language speaker and 37 million as a second language speaker [16][17]. It is not only one of the scheduled languages of India and the state language of the state of West Bengal but also the national language of Bangladesh. The Bangla language belongs to the Indo-Aryan group of languages of the Indo-Iranian branch which in turn belongs to the Indo-European language family [18]. It is believed that the Bangla language has evolved from Sanskrit and Magadhi Prakrit language [19]. Dialect-wise Bangla is divided into two main branches: Western and Eastern. The Western branch consists of Rarha, Varendra, and Kamrupa. Rarha is further sub-divided into South Western Bangla and Western Bangla. The present study is based on the Western Bangla which is the official dialect of West Bengal and popularly known as Standard Colloquial Bengali (SCB) [20]. The Bangla Language consists of 47 phonemes out of which 33 are consonants and the rest are vowels along with their nasalized counterpart.

The rest of the paper is organized as follows. A detailed proposed state phase algorithm has been explained in section 2. A brief description of the phase space parameter derived from the deviation plot along with the classification algorithm using k-NN with Euclidean distance metric and Naïve Bayes classifier is also presented in section 2. The experimental procedure for the automatic labeling of the raw signal using the state phase parameter is enumerated in section 3 followed by results and discussion in section 4. Finally, the conclusion is given in section 5.

2 Methodology

The basic objective for the bottom-up approach in speech recognition is the correct identification of the phonetic events instead of the phoneme itself. These phonetic events are realized in form of voice, noise, friction, etc. in the speech signal. For identification of the above-said signal components, a time-domain phase space algorithm is being developed for the extraction as well as identification. This section details the description of the phase space algorithm.

2.1 Phase Space

The reconstructed phase space algorithm involves the construction of the phase space portrait which will be used explicitly for capturing the dynamic trajectory of the signal. The phase portrait is constructed by plotting the time series of the signal against itself with some time lag.

Let the discrete-time series of the speech signal be represented by (1), where N represents the total samples in the discrete signal.

$$x = \{x_1, x_2, \ldots, x_N\} \qquad (1)$$

Then the reconstructed phase space (trajectory matrix) having m embedding dimensions and t delay can be expressed as in (2).

$$\begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_M \end{bmatrix} = \begin{bmatrix} x_1 & x_{1+t} \cdots & x_{1+(m-1)t} \\ x_2 & x_{2+t} \cdots & x_{2+(m-1)t} \\ . & \cdots & . \\ . & . & . \\ . & . & . \\ x_M & x_{M+2} \cdots & x_{M+(m-1)t} \end{bmatrix} \qquad (2)$$

where M= N-(m-1) *t. Here M represents the number of points on the phase portrait at the particular delay. These points are referred as phase points.

The matrix $[X_1 X_2 \cdots X_M]^T$ represents the phase points of the reconstructed phase portrait. In case of a perfectly periodic signal, it is expected that the phase points consisting of sample points having t delay corresponding to time period T or multiple of it (i.e. at a phase difference of 2π or multiple of it) will lie on the unit line as shown in Fig 1. Whereas the phase points of the aforesaid signal which have samples point at a delay corresponding to T/4 (i.e. at a phase` difference of π/2) will be most scattered from the unit line as shown in Fig 1.

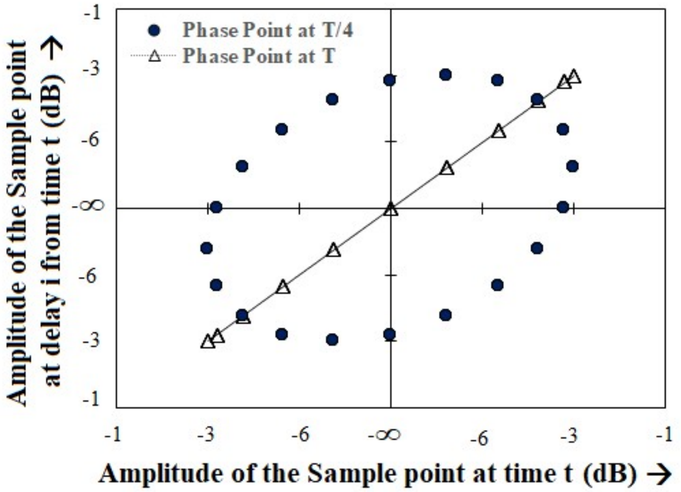

FIG. 1. PHASE PLOT OF PERFECTLY PERIODIC SIGNAL AT A PHASE DIFFERENCE OF π /2 AND 2 π

Very similar behavior can be observed in the case of the quasi-periodic signal. In this case, all such phase points having their sample points at a time delay of T will lie near the unit line and not on the line itself. This is attributed to the very fact that the source of production of the periodic signal in the case of the human being is glottis, which is a biological unit and not a mechanical unit. Hence the duration of each period in the utterance is not precisely the same. Thus, it forms a very flat and narrow region near the unit line in the phase plot as shown in Fig.2. In case of delay corresponding to T/4, the region formed by phase points in the phase space plane will have a widespread as seen in Fig. 3.

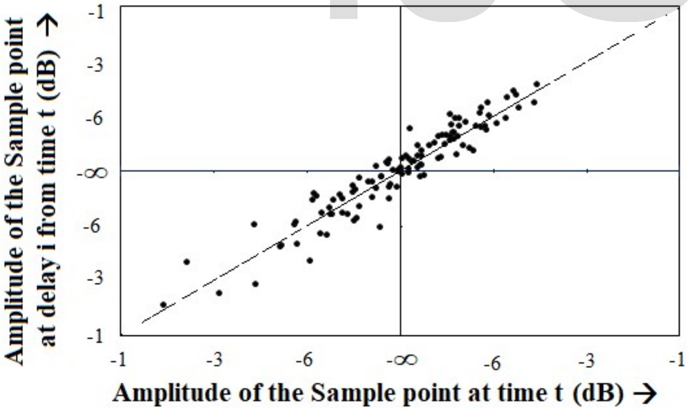

FIG. 2. PHASE PLOT OF BANGLA VOWEL /AA/ AT A PHASE DIFFERENCE OF 2 π

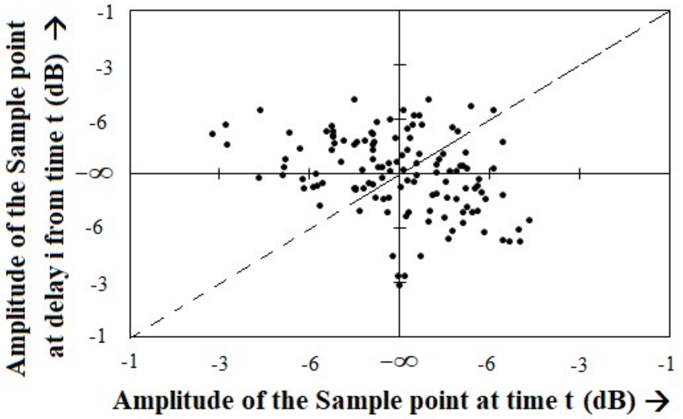

FIG. 3. PHASE PLOT OF BANGLA VOWEL /AA/ AT A PHASE DIFFERENCE OF π /4

Thus, it seems that the spread of the region in the phase portrait may be used for the determination of the fundamental frequency of the quasi-periodic signal. In order to calculate the spread or deviation of the phase point from the unit line, let us consider a phase portrait having an embedded dimension of 2. Let R (x_i, x_{i+t}) be a single phase point on the phase space plane obtained from (2) having delay t. The above said phase portrait is given in Fig. 4.

In Fig.4, let line OP be the unit line. In order to find the distance of the phase point R (x_i, x_{i+t}) from the unit (slope) line, a perpendicular is drawn from R to the unit line. Let the intersection point between R and the unit line be denoted as S. Points A and B are denoted as the intersection points between the perpendiculars drawn from point R to the corresponding axis y and x-axes respectively. Hence the coordinates of point A is [0, x_{i+t}] and B is [x_i,0]. Let the intersection point between the line RB and unit line be denoted as T. It is evident from Fig.4 that the angle between BOT, OTC, BTO, and STR is π/4. The steps for calculation of the distance RS are given in the following steps.

BT = OB = x_i [As BT and OT are the edges of the right-angled isosceles triangle OBT]

TR = RB - BT = x_{i+t} - x_i

RS2 + TS2 = TR2

2RS2 = TR2 [As TS and RS are the edges of the right-angled isosceles triangle RST]

$$RS^2 = \left(x_{i+t} - x_i\right)^2 / 2 \qquad (3)$$

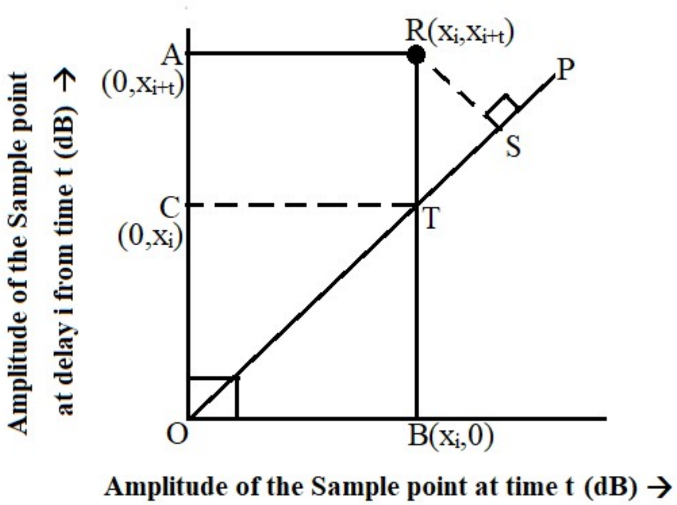
A
$(0,x_{i+t})$
$R(x_i,x_{i+t})$
P
S
C
$(0,x_i)$
T
O
$B(x_i,0)$
Amplitude of the Sample point at delay i from time t (dB) →
Amplitude of the Sample point at time t (dB) →

FIG. 4. PHASE POINT R AT T DELAY IN PHASE PLANE

Hence the deviation of a phase point from the unit line is nothing but the difference of the value of the sample points of the phase point as it is evident from (3). Therefore, the total deviation of the phase point in the phase portrait plot at a particular delay t may be defined as the sum of the square of the difference of the sample points of the phase points. This is denoted as   in (4).

$$\Delta_t = \sum (x_{i+t} - x_i)^2 \qquad (4)$$

Let Δ be the series consisting of total deviation of the phase point in phase portrait plot of the signal at various time delay t as given in (5). The plot of the above said sequence of deviation values i.e. plotted against the time delay can be referred to as deviation plot henceforth.

$$\Delta = [\Delta_1 \Delta_2 \cdots \Delta_n] \qquad (5)$$

Fig. 5b-7b gives the deviation plot of the constituent Q3 of the speech signal together with the corresponding original time-domain representation of the signal as given in Fig 5a-7a. On visual examination of all the three-deviation plots of the basic Q3 constituents of the signal reveals that there is a high chance of forming separable groups among itself. Hence the percentage of classification of the basic Q3 constituent of the speech signal will be on the higher side. The parameterization of the deviation plot will be discussed in a later section of this paper.

It is also interesting to note that in the case of the quasi-periodic segment, one of the values of the delay sequence Δ will be minimum at time period T or multiple of it (i.e. at a phase difference of 2π or multiple of it). The position of the minimum spread value in the delay sequence will give the actual time period of the signal.

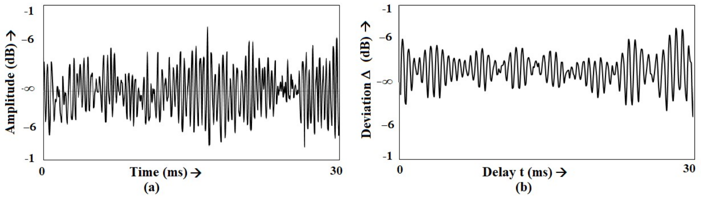

FIG. 5. (A)) QR SIGNAL AND (B) DEVIATION PLOT OF QR SIGNAL

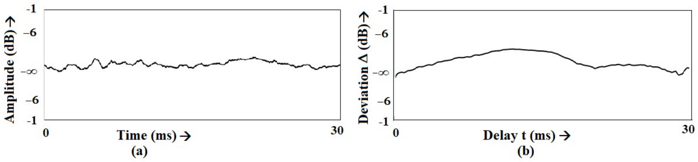

FIG. 6. (A) Q SIGNAL AND (B) DEVIATION PLOT OF Q SIGNAL

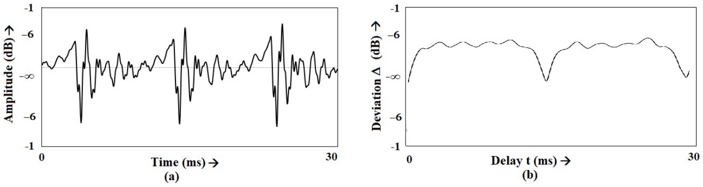

FIG. 7.(A) QP SIGNAL AND (B) DEVIATION PLOT OF QP SIGNAL

2.2 Parameters

In the previous section, we have seen that the deviation plot itself can be used to identify the basic Q3 constituent of the speech signal. Some of the following interesting observations can be made from Fig. 5-7. In the case of a quasi-random signal, the number of minima is significantly larger than that for the other two classes of the signal. A large standard deviation value/spread of the sequence of total deviation of the phase point at different delay i.e Δ is being observed for the quasi-periodic signal. Furthermore, the average value of the lowest minima in the deviation plot for the quasi-periodic signal seems to be less than that of the quasi-random signal. Finally, a flat plateau can be seen in the deviation plot for the quiescent class which is absent for the other two classes. The four parameters are formally defined as below:

2.2.1 Rate of Minima(R)

It is defined as the total number of valleys V of the sequence of total deviation of the phase points i.e. Δ over the given period of time as given in (6).

$$R = V/T \qquad (6)$$

where V is the cardinality of the set of valleys $\psi = (v_1, v_2, \ldots, v_V)$ formed the sequence of Δ and T is the duration of the window in which the feature extraction is done. The series $\{\psi_V\}$ so formed contains both the delay and the amplitude information of the deviation plot. The valleys of the series $\{\psi_V\}$ so formed should satisfy the following condition $\Delta_{t-1} \le \Delta_t \le \Delta_{t+1}$ except $\Delta_{t-1} = \Delta_t = \Delta_{t+1}$

2.2.2 Spread (δ)

Spread is defined as the standard deviation of the deviation plot that has been formed by the sequence of Δ. This is defined as in (7)

$$\delta = \left(\frac{1}{(P-1)} \sum_{t=1}^{P} \left(\Delta_t - \overline{\Delta} \right)^2 \right)^{1/2} \qquad (7)$$

2.2.3 Minimum Value (M)

The minimum value (M) of the set of valleys $\{\psi_V\}$ which are constructed from a sequence of total deviation of the phase point at different delay i.e. $\{\Delta_t\}$. Mathematical representation is given in (8)

$$M = Minimum\{\psi_V\} \qquad (8)$$

The minimum amplitude of the series $\{\psi_V\}$ is chosen in such a manner that it should lie between the admissible pitch (Φ) values. Usually Φ_{MAX} =400 Hz and Φ_{MIN} =50 Hz.

2.2.4 Flat Count (F_c)

It is defined as the total number of consecutive three points in the deviation plot which are lying at the same or near about amplitude.

$$\Delta_{t-1} \pm (\Delta_{t-1} * \delta) = \Delta_t \pm (\Delta_t * \delta) = \Delta_{t+1} \pm (\Delta_{t+1} * \delta) \qquad (9)$$

where δ is a constant and through a series of experimentation the value of the constant δ is found to be 0.05.

The length of the initial window for the extraction of the above said four parameters are set to be 15 milliseconds. Once the pitch period is detected then the length of the default window will be set to the pitch period.

2.3 Classification Algorithm

In this study, we have taken two classifiers namely the Naïve Bayes classifier and k-NN classifier using the Euclidean Distance metric for identification of the basic Q3 component of the signal. The above two different classes of classifiers have been chosen for comparison of the performance of classification of the basic Q3 component of the signal using the aforesaid four-parameter. The details of the aforesaid two classifiers are as follows.

Naïve Bayes classifier belongs to the family of a probabilistic classifier. It is basically a parametric machine learning algorithm. The main assumption of the Naïve Bayes classifier is the independence of the feature set among themselves. Let the Bayes theorem along with the above-said assumption be represented by (10).

$$P(c \mid X) = \frac{P(X \mid c)P(c)}{P(X)} \qquad (10)$$

where c represents the class label of the basic Q3 component of the signal and X represents the above-said set of four parameters, which are derived from the deviation plot. Let this X set be represented by eq 11.

$$X = (x_1, x_2, .., x_n) \qquad (11)$$

where n is the total number of parameters and j is the class number. Substituting the value of X in eq 11 then we have eq 12.

$$P(c_j \mid x_1, x_2, ...x_n) = \frac{P(x_1 \mid c_j).P(x_2 \mid c_j)....P(x_n \mid c_j)P(c_j)}{P(x_1).P(x_2)...P(x_n)} \qquad (12)$$

Equation 13 represents the generalized form of (12). Here $P(x_1).P(x_2)...P(x_n)$ is constant across all the classes and hence it is dropped from (12).

$$P(c_j \mid x_1, x_2, ...x_n) \propto P(c_j)\prod_{i=1}^{n} P(x_i \mid c_j) \qquad (13)$$

Assuming that the attribute follows a normal distribution according to the central limit theorem, the conditional probability $P(x_i \mid c_j)$ can be written as in (14).

$$P(x_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \qquad (14)$$

The class which returns the maximum probability will be assigned to the test segment. This is illustrated in (15).

$$\arg\max_c P(c_j)\prod_{i=1}^{n} P(x_i \mid c_j) \qquad (15)$$

Equation 15 will return the class with maximum probability.

The next classifier taken for the current study is k-NN with the weighted Euclidean distance as a distance metric. This classifier belongs to the class of non-parametric classifier as it does not make any assumption on the underlying data. Let μ_{ij} and σ_{ij} represents the mean and standard deviation respectively of the jth class of the ith parameter. Then weighted Euclidean distance D_j for the jth class is given in (16)

$$D_j = \sum_{i=1}^{p} \left((x_i - \mu_{ij})^2 / \sigma_{ij}^2 \right)^{0.5} \qquad (16)$$

where p is the total number of parameters. The jth class which has a minimum Euclidean distance D_j will be assigned to the test variable.

3. Experimental Details

A total of around 400 sentences uttered by 8 speakers have been taken for the current study. The selection of the speakers has been done in such a manner that it covers almost the entire age range starting from child to old of both sexes. This has been done consciously to capture any variation of the signal across the age groups and sex. The metadata of the speakers is given in Table 1.

TABLE 1
Metadata of Speaker

Speaker ID	Sex	Age Group	Nativity
Sp1	M	10-20	SCB
Sp2	F	10-20	SCB
Sp3	M	20-30	SCB
Sp4	F	20-30	SCB
Sp5	M	30-40	SCB
Sp6	F	30-40	SCB
Sp7	M	40-60	SCB
Sp8	F	40-60	SCB

In Table 1 the nativity of the informants refers to the residence and birthplace of the speaker, with both their parents belonging to Kolkata and neighborhood [21].

The average duration of each of the recorded sentences is around 4 seconds. The database of around 1000 segments for each of the basic Q3 constituents of the speech signal is carefully manually segmented from the recording. The average duration of such segments is around 50 milliseconds. The prepared dataset is equally divided into two halves; one for training and the other for testing. All the recordings have been done at the studio environment. All the speech files are recorded in a mono channel, using the sampling frequency of 8000 Hz with 16 bits encoding. Internationally for voice communication, every channel is allotted bandwidth of 4000Hz including the guard band. Accordingly sampling frequency of 8000 Hz is chosen to ensure the Nyquist rate. After the recording and careful preparation of the dataset including annotation by experienced linguists, the aforesaid four parameters are extracted.

In this paper, both Naïve Bayes and Euclidean distance algorithm have been used for comparison of classification of the basic Q3 constituent of the speech signal. An attempt has been taken to find the best and optimal combination of the above said four parameters for the classification of the basic Q3 constituent of the speech signal.

4 Result and Discussion

Fig. 8-11 shows the frequency distribution of the above-said parameters. It can be observed from all the frequency distribution charts that majorly there is an overlapping of two classes but the third one separates out quite significantly. For example, from Fig. 8 it is evident that the quasi-periodic class gets separated from the other two classes. Frequency distribution of count minima rate as shown in Fig. 9 indicates that there is a high chance of separation of quasi-random class from the rest of the two classes. Whereas frequency distribution of minimum amplitude as represented in Fig. 11 indicates the capability of using for separation of quiescent from the other two.

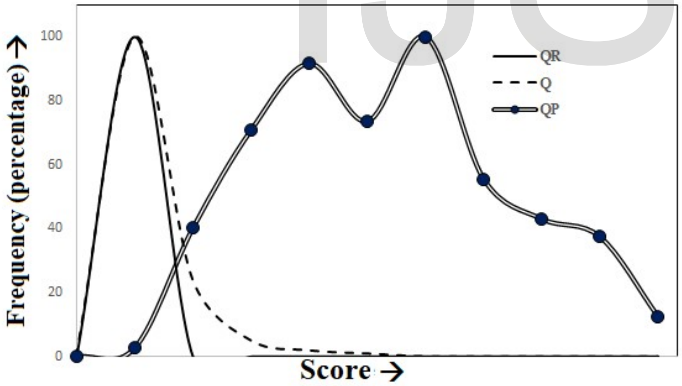

FIG. 8. FREQUENCY DISTRIBUTION OF SPREAD

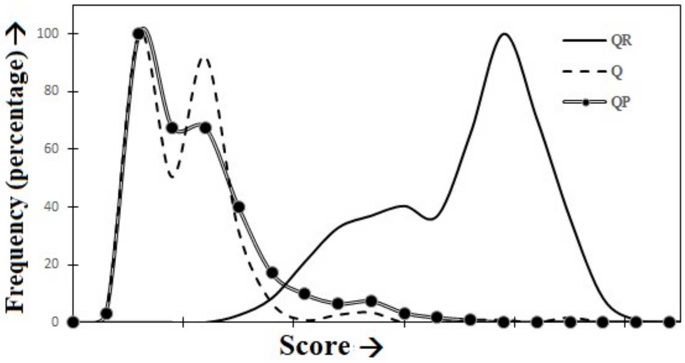

FIG. 9. FREQUENCY DISTRIBUTION OF COUNT MINIMA RATE

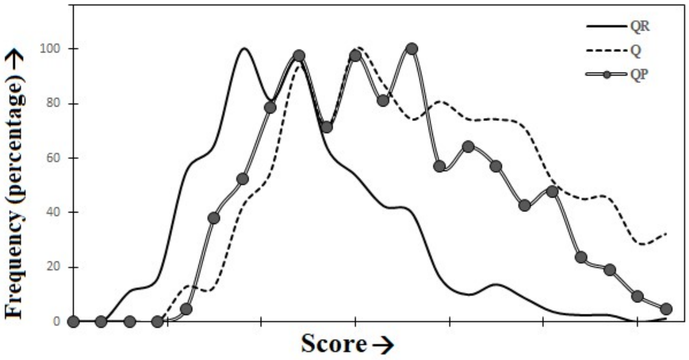

FIG. 10. FREQUENCY DISTRIBUTION OF FLAT COUNT

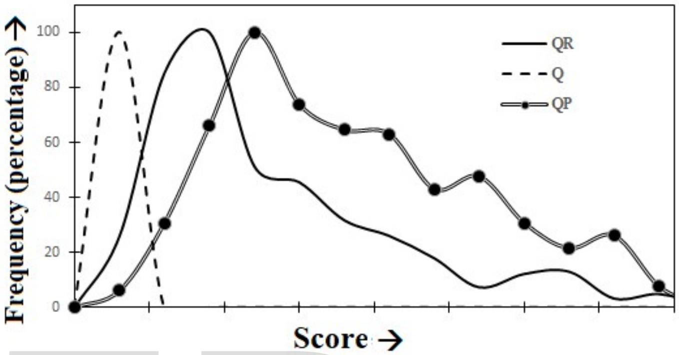

FIG. 11. FREQUENCY DISTRIBUTION OF MINIMUM AMPLITUDE

But there is a deviation of the above-said observation in the case of flat count where the basic Q3 constituent classes of the signal cannot be separated as seen in Fig. 10. Hence no single parameter may be used for classification Q3 of the speech signal. Therefore, a multi-parametric classifier is used for better classification of basic Q3 components of the signal as indicated by frequency plot of spread, count minima rate, and minimum amplitude.

TABLE 2
CONFUSION MATRIX OF Q3 FOR FOUR PARAMETERS

		Classified As		
		QR	Q	QP
Original Class	QR	**97.30**	1.09	1.28
	Q	1.47	**96.76**	1.77
	QP	2.26	0	**97.74**

The confusion matrix for the classification of basic Q3 components of the signal using the above said four parameters are given in Table 2. The overall recognition rate comes out to be 97.5 % which seems to be very encouraging. Both quasi-random and quasi-periodic have the highest rate of recognition of around 97.6%. The major misclassification of around 3% occurs for the quiescent class which gets almost equally distributed to quasi-random and quasi-periodic class. This miss classification may be attributed to the flat count parameter as indicated in Fig. 10.

TABLE 3
CONFUSION MATRIX OF Q3 FOR FOUR PARAMETERS

		Classified As		
		QR	Q	QP
Original Class	QR	**98.3**	1.0	0.07
	Q	1.4	**96.9**	1.7
	QP	2.5	0	**97.5**

Hence a three-parameter classifier excluding flat count is used for the construction of the confusion matrix given in Table 3. The overall recognition rate by using 3 parameters comes out to be 97.6%. The improvement occurs due to an increase in the recognition rate of the quasi-random signal which now turns to be 98.3%. Although we have dropped the flat count parameter from the parameter set the overall recognition set improves by only 0.1%. One of the main reasons attributed to the low recognition of the quiescent class may be the ambient noise present in the speech signal. The confusion matrix for classification of basic Q3 components of the signal is given in Table 2 and Table 3 is based on the individual period in case of quasi-periodic signal and default window size in case of quiescent and quasi-random signal for the given input segment signal.

Table 4 represents the confusion matrix for the classification of the basic Q3 component of the signal is given in terms of the maximum voting system. The description of the maximum voting system (MVS) is as follows. Each input segment consists of one or many windows of finite length. Each of these windows is then classified into one of the basic Q3 components of the signal. Based on the maximum number of occurrences of the defined class, the segment is classified to that particular class.

TABLE 4
CONFUSION MATRIX OF Q3 FOR THREE USING MVS

		Classified As		
		QR	Q	QP
Original Class	QR	**100**	0	0
	Q	0	**98.1**	1.9
	QP	1.4	0	**98.6**

It is interesting to note that the overall recognition rate has increased from 97.6% to 98.8%. Around 1.2% of misclassification occurs in a quasi-periodic and quiescent class of signal.

Table 5 represents the confusion matrix for the classification of the basic Q3 component of the signal using the Euclidean distance classifier over the same database. The overall recognition rate of the basic Q3 component comes out to be 95.5%. It is interesting to note that the major miss classification occurs in the quiescent class. Majorly around 6% of the quiescent class signal is getting confused with the quasi-periodic signal and around 2% is getting confused with the quasi-random signal.

TABLE 5
CONFUSION MATRIX OF Q3 FOR FOUR PARAMETERS USING K-NN

		Classified As		
		QR	Q	QP
Original Class	QR	**96.53**	0.55	2.92
	Q	2.65	**91.16**	6.19
	QP	2.0	0	**98.0**

It is very interesting to note that there is a noteworthy improvement of the recognition rate of identification of the basic Q3 component of the signal from 95.5% to 98.8% by using the Naïve Bayes classifier than Euclidean Distance classifier. Even though the recognition rate of 97.4% for the identification of basic Q3 components using phase space parameters has been reported in one of the studies [22] using the Euclidean Distance Classifier but still the proposed Naïve Bayes Classifier outperforms it. It is more interesting to observe that the result obtained in the previous study [22] have been done on the carefully recorded CVC non-sense syllable whereas, for the current study, the segments have been chosen from the normally spoken sentences. Thus, the improvement of the current result, obtained in the classification of the Q3 component of the signal is quite indicative of its practical usage in the speech recognition algorithm.

5. CONCLUSION

In this paper, we have explored the role of the phase space parameter in the identification of the basic Q3 component of the signal. Simple four-time domain parameters derived from the deviation plot have been used for the classification of the basic Q3 component of the signal. A speech corpus of 1000 segments of each of the basic Q3 components of the signal, is divided equally into two parts for training and testing purposes. A comparative study of the Naïve Bayes and Euclidean Distance classifier has been done for this paper. It can be seen that there is a significant improvement of the recognition rate in the classification of the basic Q3 component of the signal from 95.5% to 97.6%. It is interesting to note that there is a slight increase in the recognition rate when one of the phase space parameters, namely flat count is dropped from the parameter set. Hence the reduction in the number of parameters for the process of recognition will eventually lead to the reduction of processing time. Moreover, using the maximum voting system, the recognition rate goes up to 98.8%. Although we have taken the Bangla language for the current study but the results should be equally applicable for other languages also. Finally, we can conclude that by using the simple time-domain approach like state phase we can get a comparable or better recognition rate for classification of basic Q3 components of the signal than that of the more complex frequency domain approach [22].

REFERENCES

[1] Y. Yu, "Research on Speech Recognition Technology and Its Application," 2012 International Conference on Computer Science and Elec-

tronics Engineering, Hangzhou, 2012, pp. 306-309, DOI: 10.1109/ICCSEE.2012.359.

[2] R. Tadeusiewicz, "Speech in human system interaction," 3rd International Conference on Human System Interaction, Rzeszow, 2010, pp. 2-13, DOI: 10.1109/HSI.2010.5514597.

[3] S. M. Siniscalchi, T. Svendsen and C. Lee, "A Bottom-Up Modular Search Approach to Large Vocabulary Continuous Speech Recognition," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 4, pp. 786-797, April 2013, DOI: 10.1109/TASL.2012.2234115.

[4] J. B. Allen, "How do humans process and recognize speech?," in IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 567-577, Oct. 1994, DOI: 10.1109/89.326615.

[5] Conference: Proceedings of International Conference on Emerging Trends and Developments in Science, Management and Technology At: Delhi, India Volume: ICETDSMT-2013

[6] T. Drugman, Y. Stylianou, Y. Kida and M. Akamine, "Voice Activity Detection: Merging Source and Filter-based Information," in IEEE Signal Processing Letters, vol. 23, no. 2, pp. 252-256, Feb. 2016, DOI: 10.1109/LSP.2015.2495219.

[7] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation, Beijing, China, 1993, pp. 321-324 vol.3, DOI: 10.1109/TENCON.1993.327987.

[8] A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani and Y. Grenier, "Underdetermined Blind Separation of Nondisjoint Sources in the Time-Frequency Domain," in IEEE Transactions on Signal Processing, vol. 55, no. 3, pp. 897-907, March 2007, DOI: 10.1109/TSP.2006.888877.

[9] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time-frequency distributions", EURASIP J. Appl. Signal Process., vol. 2005, no. 17, pp. 2828-2847, 2005, DOI: https://doi.org/10.1155/ASP.2005.2828

[10] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," in IEEE Transactions on Signal Processing, vol. 52, no. 7, pp. 1830-1847, July 2004, DOI: 10.1109/TSP.2004.828896.

[11] B. Barkat and K. Abed-Meraim, "Algorithms for blind components separation and extraction from the time-frequency distribution of their mixture", EURASIP J. Appl. Signal Process., vol. 2004, no. 13, pp. 2025-2033, 2004, DOI: https://doi.org/10.1155/S1110865704404193

[12] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do and M. Hasegawa-Johnson, "Time-Frequency Networks for Audio Super-Resolution," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 646-650, DOI: 10.1109/ICASSP.2018.8462049.

[13] Suparna Panchanan, Arup Saha and Asoke Kr. Datta “AUTOMATIC SPOKEN LANGUAGE IDENTIFICATION FOR INDIAN LANGUAGES USING RELATIVE ABUNDANCE MODEL (RAM)”, FRSM, 2017, 15 – 16 December, 2017, Pages 271 – 276, Rourkela, India.

[14] C-DAC, Kolkata. (2004). Annotated Bangla Speech Corpora, www. cdackolkata . com/ html /texttospeech . htm/ corpora/fonal–product/first.htm

[15] "The World Factbook". www.cia.gov. Central Intelligence Agency. Archived from the original on 13 February 2008. Retrieved 21 February 2018.

[16] Eberhard, David M., Gary F. Simons and Charles D. Fennig ed.), 2019. “Ethnologue: Languages of the World,” Twenty-Second edition. Dallas, ISBN 978-1-55671-447-4.

[17] Bhattacharya, Tanmoy. 2000. Bangla: Bengali. In Jane Gary & Carl Rubino (eds.), Facts about the world’s languages: An encyclopedia of the world’s major languages, past and present, 65–71. New York: H.W. Wilson.

[18] Harald Hammarström ,“Linguistic diversity and language evolution,” Journal of Language Evolution, Volume 1, Issue 1, January 2016, Pages 19–29, https://doi.org/10.1093/jole/lzw002

[19] Oberlies, Thomas "Chapter Five: Aśokan Prakrit and Pāli" Archived 7 May 2016 at the Wayback Machine. In Cardona, George; Jain, Danesh. The Indo-Aryan Languages. Routledge. p. 163. ISBN 978-1-135-79711-9

[20] Bhattacharya, K., Bengali Phonetic Reader, published by Central Institute of Indian Languages, 1999.

[21] S. K. Chatterji, "Bengali Phonetics", Bulletin of the School of Oriental and African Studies, vol. 2, pp. 1-25, 1921]

[22] Chowdhury, S., Datta, A.K., and Chaudhuri, B.B. “Pitch detection algorithm using state phase analysis,” J. Acous. Soc. Ind, , vol. 28, no. 1–4, pp. 247-250, Feb. 2020 .