Grenze International Journal of Engineering and Technology, Jan Issue

Speech Recognition for Decision Making using Machine Learning Algorithms & Techniques: A Review

Devi Mampi[1] and Sarma Manoj Kumar[2]
[1]Research Scholar, Comp.Sc. Deptt., Assam Down Town University
Email: mampi_ghy2011@rediffmail.com
[2]Assistant Professor, Comp. Sc. & Engineering Deptt., Assam Down Town University
Email: manojk.sarma@adtu.in

***Abstract*—In the life of human being, communication is the key element for the evolution in various fields. Information passing to the right person following exact manner helps in both commercial and delicate stage. Now we are living in a digitized world and every means of communication is only machines such as Phone calls, text messages, talking toys, vedio games, robots etc. are information communicator. Many applications have been developed to achieve the effectual communication between two sides without any disturbances. The applications act as a mediator helping in message delivery system in different forms like: coding, text, speech signals or facial expressions. The different areas of articulatory and acoustic-based speech recognition system, conversion from speech signals to text, syllable extraction and from text to synthetic speech signals, language translation, intelligent toys, human structured robots, psychological treatment and emotion recognition; where the applications have been widely used today. Through this review paper, we have observed different analytical reports of various speech feature selection methods, various clustering techniques and coding algorithms that have been applied in voice signals to attain the stated deliverables and will focus on the supervised and unsupervised techniques.**

***Index Terms*— Speech Recognition, Assamese, Clustering, Coding Algorithm, Regression, Segmentation, Cross Talking.**

I. INTRODUCTION

Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger data sets without concern for the specific outcome [18] Prosodic features mainly estimated over the uttered speech sentence in the form of long term statistics [1]. Prosody is rich source for understanding about speech. It has great significance in emotional speech analysis also [1], [3].The clustering methods there commonly used in the segmentation or classification of medical images. For many practical issues, clustering analyses we used to explore the data structure to understand the characteristics of data. Different clustering algorithms we have proposed, including the k-means algorithm, the FCM algorithm, various improved FCM algorithms [10]-[12] and soon.

Speech technologies have been developed for decades as a typical signal processing area by the progress based on new machine learning paradigms. Speech is a vocal communication through which we achieve the information about speaker, language, message etc. Prosody is rich source for understanding about speech. It has

Grenze ID: 01.GIJET.8.1.516_1

great significance in emotional speech analysis [1], [3]. Speech recognition is the process of converting an acoustic waveform into the text form familiar to the user [2],[46].
Voice and specific utterer segmentation and clustering technology is an important front-end processing technology that evaluate changes in audio, which can further evaluate the speech processing applications along with machine language translation, emotion detection, grammar analysis and syllabification [15],[34]. Another method that is called "The endpoint detection", depends on some speech characteristics such as: time energy, zero crossing rate etc. However, these characteristics are restricted to the situation of no noise or high signal-to-noise ratio and without any result on having very low signal-to-noise ratio [35].

A. Background of the Language

Assamese is a branch of Indo-Aryan language spoken by the people of the Indian State 'Assam' in general. The people living in Assam and its adjoining states Meghalaya, Nagaland etc. term the Assamese language as the 'link' language having developed vocabulary by making use of India's national language 'Hindi' along with other northern Indian languages. The Assamese language is called as a sister of all the northern Indian languages as it has been originated from the 'Vedic vernaculars. "Indo-European family" inherited many unique phonological uniqueness that makes it unique and this is the reason that it needs a study utterly interconnected design relating to speech synthesis system (Detection & Recognition) to extract the decision [12]. In Assamese, there exist continuants of two frictionless: -- the semivowels /w, j/, four spirants /s, z, x, h/, one lateral /l/, one trill /r/ and two nasals /m, n/ which are stops along with continuants [12]. 'Asamiya', the nick name of Assamese is spoken in all the districts of Assam with varying tunes and pronunciations of the words. The people of Assam are also called as 'Assamese' in English or 'Asamiya' in local areas 'Asamiya' is the mother-tongue of the Assamese. Sanskrit, the most ancient language of the Indian sub-continent has raised The Assamese language [59] which has been accorded the status of one of the official language of Assam by the state's official language act in 1960 all over the state.
Since then, it has been serving as the lingua franca among different speech communities in the whole area.

B. Types of Speech Recognition

Speech recognition systems can be separated in several different classes by describing what types of utterances have the recognition ability. These classes are classified as follows:

- ✓ Isolated Words: It accepts single word or single utterance at a time.
- ✓ Connected Words: Linked or connected word systems (or linked utterances) have similarity with the desolated words, allowing separate utterances to indistinct with a minimal silence.
- ✓ Continuous Speech: In this recognizer users are allowed to say Continuous speeches without stop and pauses or almost naturally. During the speech operation the computer determines the content, which is the recognition task of a particular recognizer.
- ✓ Spontaneous Speech: An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features.
- ✓ Deep Speech: This is the state-of-the-art speech recognition system developed using end-to-end deep learning.

C. Syllable Structure

Structure of syllable consists of verbalization features and pronunciation that is hidden in the initial, intermediary or final position of a word [10]. The ASSAMESE phonemes syllable structure formed by vowels (V) and consonants(C) and the syllabification patterns are as follows: -

V,VC,CV,CVC,CVV,CCVC,CCV,VCC.

Rule of syllabification is to observe the different patterns of syllables that may conduct experiment with monosyllabic, disyllabic, tri syllabic and polysyllabic. And in Assamese the same rule formulates an algorithm for dividing the particular language words into syllables.

II. BASIC APPLICATIONS THROUGH SPEECH RECOGNITION FOR DECISION MAKING

It is a cognitive process or a reasoning process based on assumptions of values, preferences and beliefs of the decision maker. Decision making through speech plays a very important role in our day-to-day life like speech-to-text analysis, web-based speech recognition, speech-based music recommendation system, speech-based grammar analysis, child psychology tests in schools, intelligent toys for kids, business deals, better customer service system, emotion recognition system etc.

A. Relevant issues of ASR design

Main issues on which recognition accuracy depends have been presented in the Table I.

Table.1.

ASR TOOLS and MEDIUMS

Production	(isolated words or continuous speech read or spontaneous speech)
Speed	(slow, normal, fast)
Vocabulary or generic vocabulary;	Specific Characteristics of available training data;
ENVERONMENTS conditions	Type of noise; signal/noise ratio; working
TOOLS	Microphone; Mobile recorder
Medium	amplitude; sampling rate, distortion; echo
Speakers	Sex, Age; physical and psychical state
Speaking styles	Voice tone(quiet, normal, shouted);

B. Approaches of Speech Recognition

Basically, there exist three approaches of speech recognition. They are: -

1)Acoustic Phonetic Approach: The earliest approaches of speech recognition for finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach from "Acoustic Phonetic Approach for speech signals" by **(Hemdal and Hughes, 1967)**,[13]; where given the description existing finite, distinctive phonetic units(phonemes) in spoken language having broad characterization with a set of acoustics properties. In the long run of time and development these properties have been evaluated from the speech signal..

2)Pattern Recognition Approach: Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. In speech recognition and speech synthesis, speech segmentation is an essential basic work [36], and the quality of segmentation effect the speech recognition system. The system of overall speech signals are divided into four segments mainly and they are: -- silence segment, transition segment, voice segment, and end segment.

Among these, 'Endpoint detection' is a very important part of segmentation and the functioning is consisting of three steps such as:

(1) In the silence segment, if one of the features of short time energy or zero-crossing rate exceeds the low threshold, it is marked as the beginning of the detection speech and enter the transition segment.

(2) In the transition stage, if the energy or zero-crossing rate characteristics of consecutive frames of speech exceed the high threshold then it is said to be entering the real speech segment; otherwise, the current state is restored to the silent state.

(3) The end point detection of the speech segment can be done reversely according to the above method.

3)Artificial Intelligence Approach

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach with the ideas and concepts of Acoustic phonetic and pattern recognition methods. DNN (Deep Neural Network), RNN (Recurrent Neural Network) are the examples of artificial intelligence approach). It also includes the approach of knowledge based that utilize the linguistic, phonetic and spectrogram related information.

C. Preprocessing for Noisy Speech Recognition

Pre-processing is very important for the acoustic sound waves and it is necessary to convert it into a digital signal to achieve appropriate speech information. The most commonly used tool is the Microphone via which the acoustic wave can be converted into an analog signal and then the analog signal is passed through anti-aliasing filter for further recompense to reimbursement of any channel. The bandwidth of the signal for the Nyquist rate can be lessen by using the anti aliasing filter prior to the sampling process done. After that the analog signal is passed through an A/D converter. Now a day's A/D converters are of usually having 16 bits/sample of resolution with 8000–16000 samples/sec(second). Further, processed voice can be windowed by passing through a window (generally Hamming) along with a appropriate frame size of 10to30msec [52].

The analog speech signals are then transformed into digital signals for further dispensation. That digital signal is stimulated to the first order filters to level the signals spectrally. In this way it increases the signal's energy at a higher frequency.

D. Feature Extraction

This is another essential part to represent the speech signal by a sequence of feature vectors, where the selection of appropriate features are selected [52].

(1) Features Description: There are two basic features available for speech recognition and they are:

1) Mel Frequency Cepstral Coefficients: -MFCC extraction of sample plays the most vital role in the system of speech Signal processing and the related decision-making applications from it.
2) Pitch: - It is the quality of sound that makes it possible to judge as higher or lower on a frequency related scale[5].

In reference to the zero-crossing rate [23] A model was proposed in an earlier system and it was designed with cochlear band pass filters having nonlinear operations. In that system frequency information of the signal was achieved by zero-crossing intervals. Also, another frame work was designed for obtaining optimal forgetting factor [26].

E. Speech Databases

Speech databases are widely used in Automatic Speech Recognition having different forms. Moreover, some other significant applications are such as: Automatic speech synthesis with both dependent & independent speaker with coding and analysis via language verification, identification etc. Following rule of the acoustic environment, recording of data is done either in noise free environment like applying the sound booth or in noisy office/home environment. With different forms of the databases, some corpora are designed for the development and evaluation of speech synthesis & recognition. Besides this, other application areas like ---for instance TIMIT analysis, speaker recognition designing such as SIVA, Polycost and YOHO etc. Many databases they are recorded in one native language of recording subjects; however, there are also multi-language databases with non-native language of speakers, in which case, the language and speech recognition become the additional use of those databases. For the words of Assamese having different structure [12]. The signals can be treated as noisy signals by adding Gaussian and Babble noise with the SNR of 10dB and 15dB to create additive and convoluted noisy speech signals respectively.[16].A simple energy based utterance-end silence detection program was used to estimate the Signal to Noise ratios (SNR).An approach was designed for speech enhancement by using visually derived wiener filter[44].

III. METHODOLOGIES

From various review papers a lot of methodologies have been found and regression is one of the important methods observed [6].

A. Regression

Many regression algorithms have been tested in the research area till now to identify the best one among them. The considered algorithm is as follows.

1. Feed forward neural network with one input, two hidden and one output layers (ANN).
2. Support vector regression with linear kernel (SVR-LK)
3. Support vector regression with radial basis kernel (SVR-RBF)
   Support vector regression with radial basis kernel combined with feature forward selection (SVRRBF-FS).

The mean absolute errors for all regression techniques are reported in Table II.

TABLE II. ACCURACY VALUES OF SOME EMOTIONS

ANGER	HAPPINESS	NEUTRAL	SAD	SURPRISE
5.42	3.67	3.72	2.88	3.12
2.03	1.76	2.54	1.89	1.90
1.95	1.78	2.57	1.98	1.88
1.14	0.92	2.05	.92	1.09

In this table the maximum possible error for each emotion is 10.In the primary step the feature selection section, the mRMR algorithms is applied resulting in 200 features, these features are subsequently filtered by using a Forward Selection, the selection criteria is the Mean Absolute Error(MAE) defined as:

$$E_{mae} = \Sigma c \; \frac{\Sigma^{N-1} (S^i - V)}{MN \quad c=1 \quad i=0} \quad \text{(i)}$$

Where M is the number of classes (emotions), M=5.N is the number of patterns (speech stimuli) used for validation,N=280. Y is the output of SVR, $y = \{y_1, y_2, \ldots, y_M\}$

y C %M.

Here, S is the vector in the subjective evaluation that results in the i^{th} stimulus of $sC\#^M$ and the forward selection resulted in 43 features having the minimized MAE.

B. Two-Dimensional Mapping

When the optimal features identification is over then the regression provided by using regression function C. That function consists of the feature vector F that is extracted from the input speech signal and then only optimal features are selected via the forward selection algorithm.

C. Principal Component Analysis

Features are extracted using PCA technique accounts for a maximal amount of total variance in the observed variables. The first principal component identified as the most of the variance in the data. The second component identified accounts for the second largest amount of variance in the data without any connection with the first principal component and so on.

D. Feature Selection and Extraction

The feature selection method based on the regression algorithm was used in earlier research. In most of the regression techniques the minimum Redundancy Maximum Relevance (mRMR) algorithm has been applied to reduce the number of features from 21522 to 200. The performance achieved by mRMR algorithm was very good compared to similar filtering techniques. Generally, this algorithm has been used before applying wrapper methods to select features having a large number of irrelevant and redundant features with filtration.
A survey has been focused on feature extraction methods and the features used are the pitch, the formants, the short-term energy, the MFCCs, the cross-section areas and the Teager energy operator-based features [21].

|. The complete list of features considered are:

1. Mel Frequency Cepstral Coefficients →MFCC
2. Human Factor Cepstral Coefficients →HFCC
3. Linear Frequency Cepstral Coefficients →LFCC
4. MEL Bank Spectral Coefficients →MELBS
5. Human Factor Bank Spectral Coefficients →HFBS
6. Linear Frequency Bank Spectral Coefficients →LFBS
7. Perceptual Linear Predictive 1 →PLP1
8. Perceptual Linear Predictive 2 →PLP2
9. Perceptual Linear Predictive 3 →PLP3
10. Perceptual Linear Predictive 4 →PLP4
11. 4Hz Modulation Energy →4HzME 1
12. Mel Spectrum Modulation Energy →MSME 1
13. Fundamental frequency →F0
14. Formant frequencies →F1
15. Formant Bandwidths →Bx
16. Harmonicity →H
17. Temporal Energy →TE
18. Teager Energy Operator →TEO
19. Zero Crossing Ratios →ZCR

Already computed some of the high-level features in earlier research are: Basic characteristics, Positional characteristics, Relative characteristics, moments, Regression characteristics, percentiles. The calculation of the total number of features extracted from each speech utterance has been computed by the following formula.

$$£_{all} = £_{coef} £_{high} £_{wave} = 211 \cdot 34 \cdot 3 = 21522 \text{ --- (ii)}$$

Where $£_{coef}$ is the total number of feature coefficients, $£_{high}$ is the number of high-level characteristics and $£_{wave}$ is the numberof feature waveforms.

E. Clustering Techniques

Cluster analysis or clustering is the task of grouping set of objects in the same group are more similar to each other than to those in other groups. Clustering can be classified into two categories: soft Clustering (Overlapping Clustering) and hard Clustering (or Exclusive Clustering). For soft clustering technique, concept of fuzzy membership functions are used to cluster data, so that each point may belongs to, minimum two or more clusters with different degrees of membership value. Moreover, data will be associated to an appropriate membership value. In many situations, fuzzy clustering [16],[40],[41]is more natural than hard clustering. A new two-level segmental clustering approach was designed [24] that combined the decision tree-based state tying with agglomerative clustering of rare acoustic phonetic events.

(1) Fuzzy Rules Generation [40] --The input to our fuzzy logic [42]is energy entropy (E)short time energy (S) and zero crossing rate (Z) and the output obtained from the fuzzy is the percentage of various male and female features which are present in the given speech signal.

Both the input and output variables are fuzzified into three various sets namely; large, medium and small and male, female, male and female respectively.

(2) K-MEANS CLUSTERING TECHNIQUE

The k-means algorithm is one of the simplest unsupervised and hard clustering algorithms, which is used to classify a given data set in to various clusters. k-Means has been used to cluster time series data to get better clustering results regarding speed, simplicity, implementation impact and the possibility to assign the desired number of clusters. It produces poorresultsfortheheterogeneousdatasetduetotheinefficiencyinhandlingspeechvariability'spresentinthespeechdata.

(3) FUZZY CMEANS CLUSTERING TECHNIQUE(FCM)

The most well-known fuzzy clustering algorithm is Fuzzy C Means [37] clustering technique. It was introduced by Bezdek by modifying the crisp clustering function who introduced the idea of a fuzzification parameter (m) whose value ranges between[1,n] where n=2,that determines the degree of fuzziness in the clusters.

(4) KERNEL FUZZY C-MEANS CLUSTERING TECHNIQUE (K-FCM)

Chen and Zhang enhanced this technique by replacing the Euclidean distance with a new kernel induced distance function. This technique avoids the limitation of data intrinsic shape dependency and it's search for the initial values of the algorithm. This improves the algorithm robustness [38].

(5) KERNEL VALIDITY MEASURES

To identify the suitability of the clusters, kernel-based validity measures are used. Generally, these measures are classified into two classes :(a) The classes using membership value (b)classes using geometrical characteristics. In their study they have computed the recognition accuracies for various clustering techniques considering the time taken by each function for clustering the data. The parameters used to evaluate the FCM and KFCM clustering techniques are :i)the number of clusters ii) fuzzifier iii)stopping criterion.

An intrusion detection system was designed in energy efficient sensor network using neuro – fuzzy approach [43].

F. Feature Selection Algorithms

For designing competent algorithms for selecting a less set of features, a fast mRMR feature selection scheme had been proposed [47]. Selecting the Candidate Feature is Set to compute the cross validation & classification error for a large set of features and thus evaluated a relatively stable range of small errors [17]. The optimal number of features (denoted as n) for the candidate set is determined within the algorithm. The later type of approach like mRMR and Max-Relevance, that are also termed as "filter". The filters often select features by testing whether some preset conditions about the features and the target class are satisfied. From the practical view, the filter approach has much lower complexity than wrappers. Thus the selected features often yield comparable classification errors for different classifiers; as such features often form intrinsic clusters in the respective subspace. By using mRMR feature selection in the first-stage, they intend to find a small set of candidate features, where the wrappers can be applied at a much lower cost in the second stage.

The Fig.1 depicted the results showing the time cost for Max Deep to select a single feature is a polynomial function of the number of features, on the other hand, in mRMR, it is unchanged. For example, in NCI data; Max Deep takes about 20 & 60 seconds to select the 20th & 40th features, respectively. On the other hand, mRMR always takes about 2seconds for any feature selection.

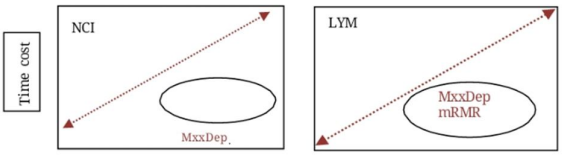


Fig.1: Feature Index

G. Features of Assamese Language and Syllables

There are 41 languages in North-East India only,[3] where Assamese is the one of the most spoken language of North East. Assamese is an Eastern Indo-Aryan language [18]. In Assamese language Labio dental, Dental, Retro flex, Uvular, Pharyngeal, Trill, Lateral fricative, phonemes are not present.

(H) Clustering for Phonemes: To cluster the phonemes, a feature extraction method had been used that is divided into two techniques temporal analysis and spectral analysis [19]. The basic feature types used for the same are Linear Predictive Code and Principal Component Analysis features, which having short term as LPC & PCA respectively of the recorded speech samples. is formed using LPC and PCA features are then used as a set to form a hybrid feature set. Pre-emphasis, Normalization, Frame Blocking, and Windowing are some of the operations, used in pre-processing system that are applied before feature extraction.

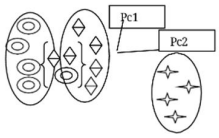


Fig.2.Dimension reduction using PCA

I. Spectral Feature Extraction

The most essential phase of speech recognition includes the tracking ability of the variation within the input speech sample. A system that can track the temporal variations of speech is more preferred to give higher successful recognition rates. Two individual parts were involved in the procedure to deal with two classes of inputs named as clusters of male & female. A Generalized Feed Forward Artificial Neural Network acted on the first block, that mapped on a class network divided the inputs into two genders based [40] clusters. A baseline speech recognition system was developed [20] with Multilayer perceptron technique to recognize the Assamese language phonemes . Moreover, some other technique Self Organized Map had been used to reduce the feature vector, .

J. Self-Organizing Map (SOM)

A Self-Organizing Map is a Neural Network architecture which can robotically produce self-organization properties during an unsupervised learning process [40],[42] where one real layer of neurons were involved. The SOM is arranged in a 2-D lattice. This architecture implements similarity measure using Euclidean distance measurement. Actually, it measures the cosine of the angle between normalized input and weight vectors having the objective is the formation of the feature map that captures the essential characteristics of the dimensional input data and maps them on the typically 2-D feature space in neural networks. The learning algorithm focuses on two essential characteristics of the map formation between neurons of the output lattice; they are competition and cooperation respectively.

K. Multi-Layer Perceptron Based Phoneme Recognizer

In our study we focused on Multilayer Perceptron that was used to design the speech recognition technique to recognize the phonemes of Assamese languages. The MLP consist of input, output and three hidden layers. A modified version of well-known Back Propagation Algorithm has been used to train the MLP in addition of avoiding the oscillations at the local minimum momentum constant that provides optimization in the weight updating process. The baseline speech recognition process consists of the following steps:

It had been observed that RNN to be a special structure of supervised ANN which suits in data recurrency technique. And the technique was trained with the clustered data to learn the number of cluster so that it can classify any unknown clustered word according to the number of cluster and issued for capturing the temporal

nature of speech signals [36]. The decision-making process of the RNN has been described in Figure3 as given below.

(a) Digitizing the speech that is to be recognized
(b) Compute the features of the speech signal
(c) Reduce the feature set using Self-Organized Map (SOM)
(d) An MLP based phoneme classifier is used to classify each set of feature corresponding to the phoneme utterance to corresponding phoneme.

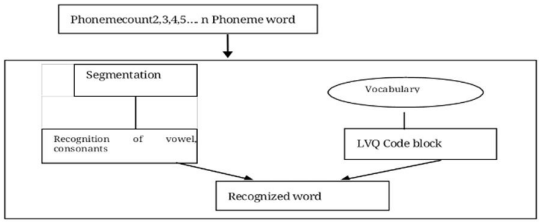


Fig.3.Phoneme Recognition Technique

The decision-making method as mentioned above had been used in case of clustered data set obtained from both SOM and KMC based clustering techniques. After computation the success rate of correct decision has been compared for both the techniques and it has been observed that the technique with SOM can provide more correct outputs (around 7% improvement) and hence it was applied on the 3-phoneme word recognition technique. To **s**elect appropriate features from a particular speech signal is an important issue to achieve high accuracy in the area of speech recognition. The term 'Feature Extraction' is the analysis of speech signal with different attributes having one hidden layer with tan-sigmoid activation function along with one input & one output layer and each layer containing log-sigmoid activation functions in it. The classification is done in terms of two classes male and female. The decision of this network is placed as a class code into the input sample vector and then passed on to two LVQ-blocks formed by SOM-MLP ANNs.

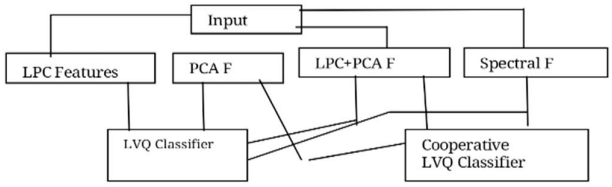


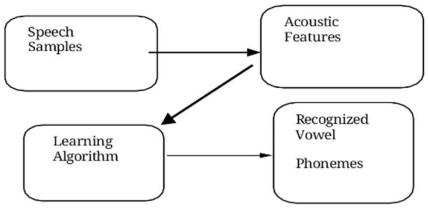
Fig.4.Block diagram of The Complete Work

L. Recognition Of The Assamese Vowels Using KNN And RNN Based Algorithm

In this system the concerned author proceeded with the steps as given below :-

1. Recording the CVC speech samples, 3speakers each from the four major dialects of Assamese language.
2. Creation of the feature vectors for training and testing purpose for the KNN and RNN algorithm.
3.Training the algorithms using the training vectors.
4.Testing for recognition of the vowels using those samples that are not used for training.
5. Comparison of the results of both the algorithms.

In this paper, recognition of the Assamese vowels and dialectal variations the acoustic features were considered for training and testing purpose. The algorithm "Scaled Conjugate Gradient" was used for the Recurrent Neural Network that gave an overall recognition rate of 84 %, while the K-Nearest Neighbor algorithm gave a recognition rate of 97 % that was better than that of the RNN algorithm. Figure 5 shows the classification output of the KNN algorithm where the legends represents the eight Assamese vowels with IPA (International Phonetic Alphabets) as/a/,/o/,/u/,/ɛ/,/e/,/i/,/ɔ/and/p/ respectively. Later, the comparison between the recognition rate and the training time of the two training algorithms was used for recognition of the vowels using the RNN and KNN.Table3 shows the individual recognition rate of the eight vowels[59].



Vowels	Recognition Rates (%)
/i/	93.33
/e/	93.33
/ɛ/	100
/a/	100
/ɔ/	93.33
/p/	100
/o/	100
/u/	93.33

The system model of the proposed work is shown in Figure5 above.

M. Cross-Language Speech Impairment Detection

Multi-language speech datasets are complex and often have small sample sizes in the medical domain. The cross-linguistic transfer methods and by collecting the large longitudinal dataset of impaired speech [32].The cross-lingual method demonstrates improvements in Aphasia detection over unilingual baselines, and the early results on the newly collected dataset show the promise to achieve a strong baseline in Alzheimer's disease detection. The translation linguistic features from speech in different languages to English using domain adaptation with Optimal Transport using a large unaligned multi-lingual dataset. The OT process minimizes the moving cost of samples from source to target using probability distribution functions. In the process mappings from other languages to English and detecting Aphasia from linguistics characteristics of speech data was much easier than some older methods.

N. Waveform Coding Algorithm

A novel waveform coding algorithm was designed based on the forward adaptive technique that performed frame by frame analysis of the input signal [48]. They used SQNR Signal to quantization noise ratio with time varying characteristics. This algorithm provided a high-speed coding to decrease the SQNR that was used for high quality transmission. The backward adaptive technique provides SQNR within 1db of the forward adaptive technique. Moreover, log uniform scalar quantizer was also implemented as it can provide higher SQNR. The proposed algorithm was applied in voice coding over the G.711 standard and comparing wide variance range with the Laplacian source, it was found that with 1 bit/sample compression provides more constant and higher level. Besides it, establishing of the quantizer asymptotic average SQNR the theoretical results were within 0.5% (with bit rate R_{Ng} >3) that pointed for achieving G.712 standard within the variance of range.

TABLE IV. THE CODING ALGORITHM: COMPARING THEORETICAL AND SIMULATION RESULTS

R_{Ng}	R	Frame length	SQNR(Dynamic)	Theory ($SQNR^{p}a$)	Lemma2($SQNR^{p}$)	Simulation ($SQNR^{p}$)
3	7.172	24	2.57	37.42	38.35	33.96
4	7.136	37	.44	37.46	38.66	35.72
5	7.115	49	.08	37.33	37.37	37.94

IV. SYSTEM MODEL FOR THE PROPOSED APPROACH:

After viewing and observing the above-mentioned review papers we have designed our model as follows:

V. RESULT ANALYSIS & DISCUSSION

We have observed many successful results from various of the previous survey on different techniques used so far by different authors and some results out of them are discussed here. The simulations were performed using MATLAB simulation tool [16]. The following two tables present the results based on the time taken by the clustering algorithms to cluster the homogeneous and heterogeneous convoluted noisy speech data. The parameter considered for the efficiency of the algorithm was based on the number of clusters and time taken by clustering algorithm.

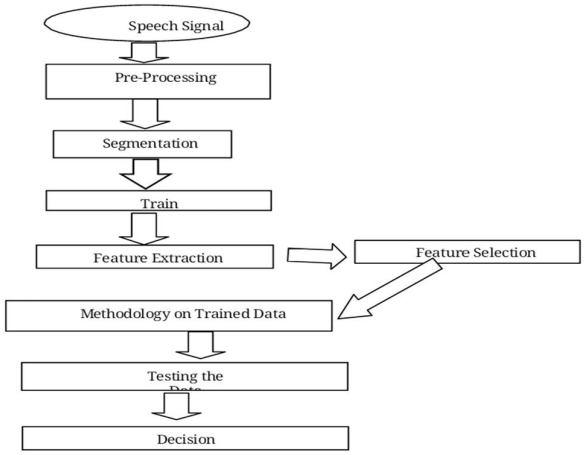


Fig.6.System Model Approach

TABLE V. EXECUTION TIME FOR HOMOGENEOUS DATA

Clusters	k-means(s)	FCM(s)	KFCM(s)
5	0.127	0.035	0.014
10	0.168	0.115	0.014
15	0.15	0.126	0.014
20	0.153	0.152	0.019

And the same for Execution time for heterogeneous data found are:0.151,0.159,0.179,0.243;0.058,0.103,0.167,0.24 ;0.014,0.014,0.02,0.02.

From the figures it seems that the as number of clusters increases the execution time taken by KFCM technique is less than FCM and k-means techniques. Considering the recognition accuracy and the type of clustering technique, the sigma value for clean and noisy speech is initialized as 0.26 and10 respectively besides this, KFCM and FCM performs relatively same for homogeneous, heterogeneous, clean, additive noise and convoluted noise for male and female speakers. Considering heterogeneous data with 10dB of SNR KFCM performs better than FCM and k-Means. Average performance of clustering algorithms with execution time for k-means, FCM and KFCM with accuracy 56, 62, 64 and execution time .1495, .099,.015 found respectively. In the analysis part, speech samples of both male and female speakers are combined to explore the suitability applicability and the performance of all the three techniques.

i) Observations ii) Clean speech signal iii) Noisy speech signals.

Generally, for homogeneous data FCM and KFCM performed better than k-means having the sigma value 0.26. Using additive noise all the clustering techniques performed equivalently good yielding 50% accuracy, whereas for heterogeneous convoluted noisy speech data the performance was reduced 5% ,yielding 50% of recognition accuracy compared to additive noise, whereas for heterogeneous data, recognition accuracies to 2-5% is observed by KFCM technique. Hence KFCM better for all types of data set. A model was designed [30] had used an application of the fuzzy expectation-maximization (EM) algorithm had used to the Baum-Welch algorithm in the HMM [49]. In HMM-based speech recognition, FVQ makes a soft decision about which code word (mean vector) is nearest to the input vector, and generates an output vector having components indicating the relative closeness of each code word to the input approach is capable of achieving higher recognition accuracy.Neuro fuzzy is represented as three layered feed forward neural network. Fuzzy sets are converted to connection weights. The Neuro-Fuzzy algorithm to adapt for human sound as this algorithm has high accuracy for classifying infant sound when compare with other popular algorithm in classification was used [45]. During transcription it seemed different problems and some interesting observation related to consonants and vowels [18]. The pronunciation rules were different from different regions of Assam. During transcription it has been

observed that there is a p occurs after every consonant. But some words like ঙ, ,ংক©, ংক do not follow this rule and in some consonant both sequences happened along with many other problems.
On the other hand, for the noisy channel (HF), the noise compensation approach leads to a significant error reduction compared to the baseline system, where more than 60% reduction of the word error rate is observed, from 27.1%downto about10% [26]. Training for a large continuous speech database using a low cost recognizer for better word error rate. In comparison with an optimized system trained with the manually generated transcriptions of the complete 72h training corpus, the word error rates relative error rate increase by14.3% and 18.6% respectively [27],[28]. Besides these mRMR features had computed to a smaller classification error, the decrement of error might not be significant for each additional feature or may be fluctuation of classification errors. Combining both Max Relevance and Min-Redundancy criteria, the mRMR incremental selection scheme provides a better way to maximize the dependency [17].The average results found for 50 samples of each of the 10 sets of noise-free, noise mixed, stress free and stressed samples were applied to SOM and MLP blocks and then the output achieved as class-code known to the trained MLP [19].Variation of the average training time of a LVQ block for 4 sl no. s for epoch 250, 500, 750 and 1000 were 81.1, 104.2, 124.1 and164.2 respectively. Whereas, variation of the average training time with lpc-vq length of the samples applied to the composite LVQ block are78.9, 132.3, 203.1, 254.2 respectively with success rates of 90.3,91.8, 93.2 & 94.8.

TABLE VI. AVERAGE RATES IN %GENERATED BY DIFFERENT FEATURE TYPES WHEN APPLIED TO A LVQ BLOCK [19]

Sl NO	Input Sample	LPC	PCA	Hybrid	Spectral
1	Male- noise-less	94.1	94.4	95.2	96.2
2	Male- noise-mixed	92.7	93.1	94.1	95.3
3	Male- stressed	92.6	92.8	93.5	95.1
4	Male- stress free	94.1	94.4	95.2	96.1
5	Female- noise less	92.3	92.6	93.5	96.1
6	Female- noise-mixed	92.4	92.8	93.5	95.6
7	Female- stressed	93.1	93.3	93.8	95.5
8	Female- stress free	94.5	95.1	95.7	96.5

Different feature types generated “Average recognition rates” were then applied to the cooperative lvq architecture for the above-mentioned inputs observed were 95.2,96.2,97.1,98.6;93.8,94.7,95.3,97.2;93.5,94.8,95.4,97.2;95.2,96.3,97.1,97.8;94.1,96.1,96.8,97.2;93.6,95.1,95.5,96.5;94.3,95.3,95.8,96.3;95.6,96.3,97.1,98.7 respectively. Some distinctive words of 5000 were listed as most frequently occurring words that had been tested by the algorithm with the function syllabify and from that almost 20,655 syllables were obtained. The algorithm achieved an overall accuracy of 98.05% on comparing with the same words manually syllabified by an expert individual**.** A syllabification rule followed an assumption by considering diphthongs as a single vowel unit [25]. A Word List of type CVC had been selected for Spoken Word Recognition Model [12], where three phases were used for experimental speech samples recording. A Model of frame level prediction and transforming into sequential prediction for Speech Recognition System as Mathematical Expression as below [29].

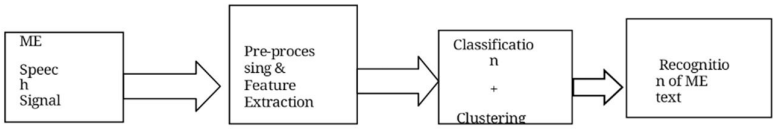


Fig.7. Mathematical Expression of SR

The RNN is trained to estimate the number of clusters residing in the data. If RNN cannot classify the data into any of the defined class, then the word is clustered with KMC algorithm fork-value 4 and the process is repeated. If the RNN again fails to classify the data then the word is clustered with k value 5 and the same process is repeated again. Thus, the proposed logic is used to determine the possible number of phonemes in the word for 3-, 4- and 5-phoneme words. Early results on renowned People show accuracy of 73.33%, sensitivity of 71.25% and specificity of 75% at subject-level, where the data set includes samples from 1-5years before diagnosis and healthy data[32].The baseline speech recognition process for the Assamese phonemes with Self Organized Map

to reduce the feature vector consisted of the three steps as mentioned above. Speech data from the same group of speakers were used with both clean and different level of noise data to train the proposed system as an experiment. Recognition accuracy obtained using clean speech was 100% for LPCC and MFCC and for the same & different speakers it was 94.23% and 89.14% respectively. The same for noisy speech obtained with the feature set of LPCC & MFCC for SNR 20 dB, 15 dB &10 dB were 73.27%, 97.03%, 59.41%, 85.15%, 47.52% & 68.32% respectively.

VI. Conclusion & Future Scope

After reviewing various techniques and algorithms we can conclude that a good way is to consider plotting the identified outlier values with some models based on regression and clustering technique, the slopes and intercepts, parameters are estimated. So, the presence of non-informative variables can lead to uncertainty of the predictions and thus reduce the overall effectiveness of the model. The supervised and unsupervised methods of feature selection are also important. Feature selection is also related to dimensionally reduction techniques in some methods seek fewer input variables to a predictive model. For unlabeled data unsupervised methods have been found more efficient than the supervised techniques. We conclude that, as speech data are mostly unlabeled in terms of feature so unsupervised techniques will suite more. Filter feature selection methods having statistical techniques to evaluate the relationship between each input variable and the target variable with some scores that can be used to choose those input variables effective in the model.
Finally, some machine learning algorithms that will perform feature selection automatically as part of learning the model as intrinsic feature selection methods having predictors that help to maximize the accuracy. During these procedures the model can pick and choose the representation of the data that suits the best.

Acknowledgement

This paper entitled “Speech Recognition Using Clustering for Decision Making: A Survey” would not have been possible
Without the support and encouragement of many people who directly or indirectly offer their help and cooperation towards the completion of the work.
Firstly, I would like to express my sincere gratitude and heartfelt thanks to my guide Dr. Manoj Kr. Sarma. I am very fortunate for having an opportunity to work with him from which I have been benefited enormously.
Last but not the least, I offer my sincere gratefulness to my parents, whose support and inspiration this work would not have been possible.

References

[1] Jakovljevi Niksa et al. “Speech Technology Progress Based on New Machine Learning Paradigm” Hindawi Computational Intelligence and Neuroscience Volume 2019, Article ID 4368036,25June201919pages.
[2] Devi Mousmita et al. “A study on issues of Speech recognition system” Journal Of Harmonized Research in Engineering ISSN 2347–7393, 2014, P246-250.
[3] Davletcharova Assel et al. “Detection and Analysis of Emotion from Speech Signals” Procedia Computer Science Volume58,2015,Pages91-96.
[4] Lugovic S.et al. “Techniques and Applications of Emotion Recognition in Speech” in 39th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO May,2016,P1551-1556.
[5] Kalita Bimal Kumar et al. “A MODEL & DESIGN OF A DATABASE OF FUNCTIONAL AND EMOTIONAL INTONATION WITH REFERENCE TO ASSAMESE LANAGUAGE” published in Journal Of Harmonized Research in Engineering ISSN2347–7393V2[4],2014,P372-376.
[6] Hicham Atassi et al. “Emotion Recognition from Spontaneous Slavic Speech”, in Cog Info Com 2012 3rd IEEE International Conference on Cognitive Info communications, December 25, 2012 P389-394.
[7] Basharirad Babak et al.“SpeechEmotionRecognitionMethods:ALiteratureReview”inTHE2NDINTERNATIONALCONFERENCEONAPPLIEDSCIENCEANDTECHNOLOGY2017(ICAST’17)October2017,V020105,P1-7.
[8] Deokar Priya et al. “ANALYSIS OF SPEECH RECOGNITION METHOD MFCC AND LDA CLASSIFER FOR HINDI WORDS” in International Journal of Scientific Research and Review ISSN No.: 2279-543X Volume 07,Issue 05, May 2019 UGC Journal No.:64650 P1780-1788.
[9] Raheel Aasim et al. “Physiological Sensors Based Emotion Recognition While Experiencing Tactile Enhanced Multimedia” Jul, 2020, 213390/s20144037, PMCID: PMC7411620PMID:32708056V4037,P1-19.
[10] Michael G.R.et al. “Comparative Study Of Various Classifiers Using Assamese Speech Utterances” in International Journal of Computational Engineering Research (IJCER) ISSN (e): 2250 – 3005 || Volume, 08 ||Issue,6||Jun–2018||P3-

38.
[11] James H. Immanuel et al. “EMOTION BASED MUSIC RECOMMENDATION SYSTEM” in International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 March 2019, Volume 06, Issue03, P2096-2099.
[12] SARMA MOUSMITA et al. “Recognition of Assamese Spoken Words using a Hybrid Neural Framework and Clustering Aided Apriori Knowledge” in WSEAS Transactions on Systems July 2013, Issue 7, Volume 12,P360-370.
[13] Dixit Ranu et al. “Speech Recognition Using Stochastic Approach: A Review” referring “Acoustic Phonetic Approach for speech signals” by(Hemdal and Hughes 1967)in International Journal of Innovative Research in Science, Engineering and Technology, February2013Vol.2,Issue2.
[14] Trivedi Ayushi et al. “Speech to text and text to speech recognition systems-A review” in IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Vol. 20, Issue 2, Ver. I (Mar. - Apr.2018),PP36-43.
[15] Jiang Nan et al. Research Article “An Improved Speech Segmentation and Clustering Algorithm Based on SOM and K-Means” in Hindawi Mathematical Problems in Engineering, 27July 2020, Vol2020, Article ID 3608286, P1-19.
[16] Vani H.Y. et al. “FUZZY CLUSTERING ALGORITHMS-COMPARATIVE STUDIES FOR NOISY SPEECH SIGNALS” in ICTACT Journal on Soft Computing, 2019, Vol.9, Issue:3, P1920-1926.
[17] Long Fuhui et al. “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy” in IEEE Trans Pattern ANALYSISAND MACHINE INTELLIGENCE, August2005, vol.27, P1226-1238.
[18] Sarma Himangshu et al. “Development and Transcription of Assamese Speech Corpus” in National Conference Computation and Language(cs.CL),27September2013.
[19] Sarma Manash Pratim et al. “Assamese Numeral Speech Recognition using Multiple Features and Cooperative LVQ Architectures” in World Academy of Science, Engineering and Technology International Journal of Electronics and Communication Engineering 2011,Vol:5, No:9, P1263-73.
[20] Bhattacharjee Utpal et al. “A Comparative Study Of LPCC And MFCC Features From The Recognition Of Assamese Phonemes” in IJERT January, 2013, Vol.2 Issue1, P.1-6.
[21] Ververidis Dimitrios et al. “Emotional speech recognition: Resources, features, and methods” in Google Scholar Speech communication 48(9) April (2006) 1162-1181,P.1-20.
[22] Liu Yang et al. “Enriching Speech Recognition with Automatic Detection of sentence Boundaries and disfluencies”, IEEE Transactions on Audio, Speech and Language processing, July 2006, V.14, No.4, P.1-15.
[23] Kim Doh-Suk et al. “Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments”, in IEEE Transactions On Speech And Audio Processing, January 1999, Vol. 7, No. 1,P55-69.
[24] Reichl Wolfgang et al. “Robust Decision Tree State Tying for Continuous Speech Recognition” in IEEE Transactions on Speech and Audio Processing, September 2000, Vol.8,No.5, P555-566.
[25] Thakuria Laba Kr. Et al. “Automatic Syllabification Rules for ASSAMESE Language” in Int .Journal of Engineering Research and Applications www.ijera.comISSN: 2248-9622, February 2014, Vol.4, Issue 2(Version1), pp.446-450.
[26] Afify Mohamed et al. “Sequential Estimation With Optimal Forgetting for Robust Speech Recognition” in IEEE Transactions On Speech And Audio Processing, January 2004, Vol.12,No.1.
[27] Wessel Frank et al. “Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition” in IEEE Transactions On Speech And Audio Processing, January2005,Vol.13,No.1,P23-31.
[28] Garau Giulua et al. “Combining spectral representation for large vocabulary continuous speech recognition” in IEEE Transactions On Audio, Speech And Language Processing, Jan2008,Vol.16,No.1,P1-11.
[29] A Vaishali et al. “Speech Recognition System Approaches, Techniques and Tools For Mathematical Expressions: A Review” in INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, AUGUST 2019, ISSN 2277-8616VOLUME8, ISSUE 08,P 1255-63.
[30] Tarihi Mohammad Reza et al. “A New Method for Fuzzy Hidden Markov Models for Speech Recognition” in IEEE International Conference on Emerging Technologies, February 2005, ISSN: 1305-5313,V4,P36-40.
[31] Fraser Kathleen C. et al. Linguistic “Features identify Alzheimer’s disease in narrative speech” in Journal of Alzheimer’s Disease, 2016, V 49(2):P407–422.
[32] Novikova Jekaterina et al. “On Speech Datasets in Machine Learning for Healthcare,” in Research Gate 15 January 2020.
[33] Anusuya M.A. et al. “Speech Recognition by Machine: A Review” in (IJCSIS) 2009, vol.6, No.3, P181-205.
[34] Delacourt P.et al. “DISTBIC: A speaker-based segmentation for audio data indexing” in Elsevier Speech Communication, 2000, vol.32, no.1-2, pp.111–126.
[35] Wu Bing-Fei et al. “Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments” in IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, SEPTEMBER 2005, VOL.13,NO.5.
[36] Hannun Awni et al. “Deep speech: scaling up end-to-end speech recognition” in Computer Science, 19 Dec,2014, vol.17, pp.1–12,2014.
[37] Bora Dibya Jyoti et al. “A Comparative Study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm” in International Journal of Computer Trends and Technology, April 2014, Vol. 10, No. 4, pp. 108-113.
[38] Yadav Asmita et al. “An Improved K-Means Clustering Algorithm” in International Journal of Computing Academic Research (IJCAR)ISSN2305-9184,(April2016),Vol.5,Number2pp.88-103.
[39] Tran Dat et al. “A Fuzzy Approach to Statistical Models in Speech and Speaker Recognition” in Proceedings of

International Conference on Fuzzy Systems of IEEE ISSN: 1098-7584, August 2000, pp 1275-1280.
[40] Meena Kunjithapatham et al. “Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network” in The International Arab Journal of Information Technology, September 2013,Vol.10,No.5,PP477-485.
[41] Vani H.Y. et al. “Fuzzy Speech Recognition: A Review”, in International Journal of Computer Applications (0975–8887),March2020,Vol.177,No.47,PP39-54.
[42] Jawad Lubna et al. “Arabic Voice Recognition Using Fuzzy Logic and Neural Network”, in International Journal of Applied Engineering Research, ISSN: 0973-4562, July 2019, Vol.14, Number3, pp.651-662.
[43] Mittal M et al. “A Neuro-Fuzzy Approach for Intrusion Detection in Energy Efficient Sensor Routing” in Proceedings of the 4th International Conference on Internet of Things: Smart Innovation & Usages (IOT-SIU),Ghaziabad,18–19April2019,PP1-5.
[44] Almajai Ibrahim et al. “VISUALLY-DERIVED WIENER FILTERS FOR SPEECH ENHANCEMENT” inICASSP’07, April 2007, Vol.19, Issue6.
[45] Shah Maitri et al. “Speech Recognition using Neuro Fuzzy Network”; IJARIIE-ISSN (O)-2395-4396, Vol-3, Issue-2,PP4852-4857.
[46] G. Hemakumar et al. “Speech Recognition Technology: A Survey on Indian Languages”; in International Journal of Information Science and Intelligent System, October 2013, Vol.2, No.4.
[47] Nereveettil Catherine J.et al. “Feature Selection Algorithm for Automatic Speech Recognition Based On Fuzzy Logic” in International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3, Issue1, January2014.
[48] Peric Zoran et al. “An adaptive waveform coding algorithm and its application in speech coding” in Elsevier Digital Signal Processing, January2012, Vol.2, Issue.1, PP 199-209.
[49] Deng Li et al. published “Dynamic Compensation of HMM Variances using the Feature Enhancement Uncertainty Computed from a Parametric Model of Speech Distortion” in IEEE Transactions on Speech and Audio processing, June2005,vol.13,no.3,PP412-421.
[50] Www. google. Com as web reference.
[51] W. AL-Sawalmeh et al. “The Use of Wavelet Entropy in Conjunction with Neural Network for Arabic Vowels Recognition”, WSEAS Transactions on Signal Processing, Issue3, volume7, 2011.
[52] Laba kr. Thakuria & Prof. Pran Hari Talukdar, “Analysis and Detection of Emotional states of a person by extracting Pitch and Formants with respect to Assamese Language” at International Journal of Scientific &EngineeringResearch,Volume5,Issue2,February-2014890ISSN2229-5518IJSER©2014.
[53] Reddy, T. et al. “A deep neural networks based model for uninterrupted marine environment monitoring”.Computer.Communication.2020, 157, P64–75.
[54] Mittal, M. et al. A Neuro-Fuzzy Approach for Intrusion Detection in Energy Efficient Sensor Routing, In Proceedings of the 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India,18–19April2019.
[55] MARK W. JONES and XIANGHUA XIE, “Clustering and Classification for Time Series Data in Visual Analytics: A Survey”, published in IEEE ACCESS multidisciplinary open access journal on December 10,2019,Digital Object Identifier 10.1109/ACCESS.2019.2958551,volume7,pno:181314-181338.
[56] Vit Niennattrakul & Chotirat Ann Ratanamahatana, “ On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping” in Conference: 2007 International Conference on Multimedia and Ubiquitous Engineering(MUE 2007),26-28April2007,Seoul,Korea,DOI:10.1109/MUE.2007.165.
[57] Noorus Sabah; 4G Technology and its Applications; International Journal of Research in Engineering, Technology and Science,VolumeVI,SpecialIssue,July2016.
[58] Dr. Amin Salih Mohammed et al. “4G and 5G Communication Networks Future Analysis”, in International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume12, Number3 (2019), pp.343-349.
[59] Mridusmita Sharma et al., “Dialectal Assamese Vowel Speech Detection using Acoustic Phonetic Features, KNN and RNN”, International Conference on Signal Processing and Integrated Networks (SPIN) May, 2015, PNo48-52.
[60] PALLABI TALUKDAR et al. “Recognition of Assamese Spoken Words using a Hybrid Neural Framework and Clustering Aided Apriori Knowledge”, WSEAS TRANSACTIONS on SYSTEM, E-ISSN: 2224-2678, issue 7, Volume12, July2013, pno360-370.