Grenze International Journal of Engineering and Technology, Jan Issue

Accuracy Comparison: An Approach of Unsupervised Machine Learning Algorithm for Assamese Speech Recognition

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 recent years the Automatic speech recognition systems have made vivid acts. Still, the key to making recognition more robust is to reduce the difference between training and testing with speech that commonly held. The postulation of ASR (Automatic speech recognition) is becoming less and less feasible, because ASR applications change from firmly measured to more usual situations with a erratic number of random sound bases. Decoding the speech source of interest while listening to several sound sources at the same time seems a more accurate description of the ASR process with different features that suits these challenging environments. The aim is to compare the accuracy with different features at different instance for robust ASR into two sub-problems:(a) identification/separation of the speech and noise using speech properties alone and (b)recognition based on the comparison of the accuracy and entropy. The basic assumption is that some regions of the speech time-frequency representation remain relatively unaffected by the noise, that they can be identified and that they alone are sufficient for ASR. In contrast to conventional techniques which require models of all sources in the auditory scene and their subsequent decoding even when only one of the sources is of interest, the techniques described in this paper make no such requirement. However, they are flexible enough to use this information if itis available. In the adaptation of a conventional Hidden Markov model (HMM) based ASR system two techniques are used for partial indication: (i) down grading of the state dispersals, so that only the likelihood of the reliable sections are evaluated; and (ii)attribution of the undependable sections by replacing the unreliable features with a single point from the community restricted deliveries. In the experiments, the reliable features are identified via clustering accuracy derived through K-means clustering. The potential of the techniques is indicated by using the clean speech to identify the reliable regions in the noisy speech.**

***Index Terms*— Speech, Assamese, K-means, Accuracy, Vowels and Cluster.**

I. INTRODUCTION

Now a day's ASR systems perform very well enough in many areas of applications. Still, changing the applications from tightly controlled to real world environments remains a challenge. The resettlement detects many problems of ignorance about how to model the sources of speech unpredictability in changing auditory

Grenze ID: 01.GIJET.8.1.517_1

environment, earlier enclosed by the assumption of a calm environment and single source auditory vista. The aim of this paper is to present and discuss various features that have been used to increase the solidity of the ASR systems. Introducing some factors contributing to the speech inconsistency along with the changeability due to noise. Subsequently, it has been introduced the notion of an acoustic environment to describe the interaction between the speech and the noise. There are many ways like speech, hand motions, facial mien etc. that people communicate with each other. Among all these, speech is measured as the most important means that we use, because it simplifies communication that is the most widely used by the people. Several letters turn into several words accompanied by vocal sounds, that make a speech and that results in the best way of message passing. In objects of air or in empty and non-living form of waves, the speech can be spread. Also, in the form of small circles of the sound source or a wave that overlaps between them, which is characterized and then extends these circles gradually until they completely vanish while spreading over elongated distances. For the speakers, who talk the same language i.e., the sender and recipient have the same keys to decipher the meaning of the message then the 'Logical Speech' is used. This phenomenon has been applied by the researchers and developed it as a base in the communication between humane and the machine where the sound has helped to facilitate the use of the machine. Automatic speech recognition has developed the artificial intelligence seeking to create very flexible methods of handling the machine and allowing the user to communicate and exchange information without using known input/output units like the keyboard. Also, in several areas, such as the caring some disabled person, cars while driving, anguish calls etc., voice-based input/output techniques are very convenient.

We conclude with a summary of clustering techniques with best effective feature through this paper. However, the value of the summary in assessing the relative merits of the techniques is limited. The researchers have been using vastly different systems and speech corpora, leading to inability to compare the techniques across different research groups [1].

A. Origin of Speech Inconsistency

Some common reasons for unpredictability in speech are:(a) ineffective with noise (additive, convolution, resonance); (b) talking stylishness (Lombard consequence, talking rate); (c) inter utterer differences (speech superiority, arena, gender, talk); (d) framework (negotiation, transcription, conversation). The degradation occurring due to "cleaning" are particularly harmful as they are non–linear, unnatural and their extent is hard to quantify in advance. On the other side, the present statistical ASR systems are quite sensitive to degradation producing data unseen during the training (the problem of statistical "outliers").

B. Speech Enhancement

Techniques from this class aim to reduce the statistical difference between clean training and noisy testing features using some apriori knowledge about the speech, the noise and/or the way they combine. All systems today use some of the techniques from this class. Most of them originate from attempts to improve speech intelligibility. They all introduce a new problem to the robust ASR as well: the "boosted" or "dressed" to deal with noise reduction, speech may be more comprehensible to the humans.

Voice quality improvement is mostly striking when other ASR system apparatuses can't changed [2]. One way to limit the damage is to train the ASR system on "cleaned" clean speech so that the statistical models train to handle the degradation. Another possibility is to modify the enhancement process in such a way to balance the enhancement and the degradation of the speech. The ASR can also be defined as graphical depictions of frequencies produced as a time function. Speech fusion, dispensation and utterer credentials are involved in all speech processing techniques. To create speech lines or to make speech communication in many cases Speaker verification is very essential. Some of the applications of Speech recognition are listed below:

- Voice services: speaking clock, weather, race results, etc.,
- Quality control, data entry.
- Avionics, Training
- Disabled assistance, Vocal dictation

Now a days, in mobile phones or in cars: car radio, air-conditioning, on-board navigation on the Internet with voice commands voice recognition modules have been implanted.
Wiener filtering is commonly used as alternative or complementary technique to spectral subtraction for removing additive noise. The filter is designed to minimize the minimum mean square error in time domain. It is a maximal likelihood filter if the distributions of the speech and the noise are Gaussian [2].

C. Characteristics of Speech Recognition System

In the voice recognition systems, many variables are involved and we must be familiar with the variables. Otherwise, it's very difficult to choose appropriate algorithm for the process using the proper functioning of the variables:
Common studies injunctions some types of Speech, which are mainly of four types:

1. Isolated Words: This type usually requires a quiet (silence state) between utterances.
2. Connected Words: Word systems are similar to isolated words, the only difference between themes to allow separate words to "run-together" with a minimum of pausing between them.
3. Continuous Speech: The users of this type talk almost normally, while the computer selects the content. It is one of the most difficult systems. Speech vocabularies are as follows: --
   - Bundle of words (Small vocabulary)
   - Hundreds of words (Medium vocabulary)
   - Thousands of words (Large vocabulary)
   - Tens of thousands of words (Very-large vocabulary)
4. Dependency of Speaker
   - Systems that require the user to train the system according to the user's voice (Speaker dependent system).
   - Systems developed for any speaker (Speaker independent system).
   - Developed to adapt to the characteristics of new speakers (Speaker adaptable system).

II. Background of the Assamese Language

The Assamese language has a root at Indo- Iranian sub family. It has total three allied languages. They are---Hazong, Bisnupuria and Sakma. Hazong is used by some local people of the Goalpara district of Assam, Garo hills of Meghalaya and north border of Bangladesh. All of them have some similarity with modern Assamese. Root of the Assamese language is Indo- European family of languages. The eminent linguistic of Assam Dr. B. K. Kakoti pointed out in his Ph.D. Thesis "Assamese: Its formation and Development" that there are some words in Assamese which are imported from Indo-Chinese family of language. Bisnupriya language is spoken by people of Barak valley and its originis at Bangladesh. All Indo-European languages has at least few similarity with Assamese language. For example শ, স, ষ phonemes are found only in Assamese language, not in any other Indian languages. But many words of these three pronunciations are available in Pharusee, Germany, Scotish or Greek .They are entering to Assamese alphabet from German. Again the language bears some similarity with North-East Asian language Ainu used by the Japanese. It is the ancient language of Japan. Due to migrants from north India and for Hindu religion fast spreading in this area, the Assamese is greatly influenced by Sanskrit. Sanskrit is basically Drabir based language, so there is vast use of cerebral pronunciation. Cerebral pronunciation is not found in Assamese or European languages. Depending upon geographical variation dialect also varies from place to place and time to time. This is also noticed that in most of the Indian languages if two consonant occurs together in front of a word, pronunciation converts to one. Actually, all these are characteristics of a language influenced by geographically different areas. The ASSAMESE phonemic structure consists of eight vowels, ten diphthongs, twenty-one consonants and two semi vowels. Bisnupriya language is spoken by the people of Barak valley and its origin was at Bangladesh. Researchers say that dictionary of any language carries history of every word. In some geographical areas particular phones are never used. In Assam particularly in Barak valley /kh/ is used instead of /h/.This is also noticed that in most of the Indian languages if two consonant occurs together in front of a word, pronunciation converts to one. Actually all these are characteristics of a language influenced by geographically different areas.[13]. The ASSAMESE phonemic structure consists of eight vowels, ten diphthongs, twenty-one consonants and two semi vowels [12] and the syllabification rule of vowels & consonants leads to a new way of tracking the voice [8].

TABLE I. IPA SYMBOLS OF ASSAMESE VOWELS

VOWELS						
	Alveolar		Velar		Glottal	
	IPA	Script	IPA	Script	IPA	Script
Close	i	ই / ঈ			uû	উ/ঊ
Neau Close					ʊ	ও
Close mid	eê	এ			oô	অ'
Open mid	ɛ	এ			ɔ	অ
Open			a	আ		

III. DATABASE DESCRIPTIONS

A. Data Collection

As we have done our research for this paper during the covid pandemic, due to which travelling is being restricted in many areas. Therefore, we have collected voice recordings only from family members with a mono audio channel recorded using Praat software installed on Intel core i3 processor. Later on, we will collect recordings from other channels belonging from Assam.

The term robust features refer to applying a transformation in the first stage of processing of the speech signal (feature extraction) Auditory motivated robust features. Since human audition is so robust to noise, many researchers have tried to replicate the better known parts of the human auditory chain in hope of achieving robustness [3]. Our selected features with descriptions are listed in the Table2 below.[4].

TABLE II: FEATURES DESCRIPTION

SL No	Selected Features	Full Abbreviations	Feature's activity on Speech
1	MFCC	Mel Frequency Cepstral Coefficents (MFCCs)	MFCC are a feature widely used in automatic speech and speaker recognition. It gives a discretecosine transform (DCT) of a real logarithm of the short-term energy displayed on the Mel frequencyscale .The mel scale is a scale of pitches judged by listeners to be equal in distance one fromanother.
2	HFCC	Human Factor cepstral coefficients	HFCC is modified form of MFCC and it outperform in clean condition generally used for tracking of speech content in frequency domain.
3	LFCC	Linear Frequency Cepstral Coefficient	LFCC the analysis of changes in pre-emphasis, numbers of filter bank and numbers of cepstral are conducted.
4	MELBS	MEL Bank Spectral Coefficients	MELBS is computed by multiplying the Discrete Fourier Transform (DFT) module of a speech segment by the corresponding mel frequency filter bank.
5	HFBS	Humane Factor Bank Spectral Coefficients	HFBS is computed by multiplying the Discrete Fourier Transform (DFT) module of a speech segment by the corresponding human-factor filter bank .
6	LFBS	Linear Factor Bank Spectral Coefficients	LFBS is computed by multiplying the Discrete Fourier Transform (DFT) module of a speech segment by the corresponding Linear factor filter bank.
7	Harmonicity	Harmonic ratio	Harmonicity object represents the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio (HNR).It can be used as a measure for: The signal-to-noise ratio of anything that generates aperiodic signal,voice quality.
8	4HzME	4Hz Modulation Energy	It is the variations of an acoustic signal (both spectrally and temporally) that shape how the acoustic energy with 4Hz fluctuates as the signal evolves.
9	MSME	Mel Spectrum Modulation Energy	MSME to be used If a time-series input y, sr is provided, then its magnitude spectrogram S is first computed, and then mapped onto the mel scale by mel_f.dot(S**power). By default, power=2 operateson a power spectrum.
10	FundFreq	Fundamental Frequency	Fundamental frequency is component of a voice, which are the lowest frequencies of a periodic voice waveform.
11	Mean Freq	Mean Frequency	Mean Frequency of a spectrum is calculated as thesum of the product of the spectrogram intensity (in dB) and the frequency, divided by the total sum of spectrogram intensity.

12	Median Freq	Median Frequency	Median Frequency is the middle value in an orderedset of frequency. If there is an odd number of observations, the median is the middle number. If there is an even number of observations, the median will be the mean of the two central numbers.
13	StdDeviation	Standard Deviation	Standard Deviation is the average distance (or deviation) from the mean. The mean is the sum of the product of the midpoints and frequencies divided by the total of frequencies.
14	Formant Freq	Formant Frequency	Formant Frequencies is a concentration of acoustic energy around a particular frequency in the speechwave.
15	FB	Formant Bandwidth	Formant Bandwidths-It is the difference in frequency between the points on either side of the peak which have amplitude A/(sq. root of 2).
16	ZCR	Zero Crossings Ratio	Zero Crossing Ratio gives information about the number of zero-crossings present in a given signal.
17	PLP1	Perceptual Linear Predictive1	PLP1 is simply the Bark spectrum of a speech segment. The PLP method analyzes the onset strength envelope in the frequency domain to find alocally stable tempo for each the Bark scale is a psycho acoustical scale The Bark scale ranges from1 to 24 Barks, corresponding to the first 24 critical bands of hearing.
18	PLP2	Perceptual Linear Predictive2	Plp2 is the post- processed Bark spectrum, where PLP1 is multiplied by the equal loudness curve andthe module is compressed.
19	PLP3	Perceptual Linear Predictive3	Plp3 type is the LPC smoothed spectrum of PLP2.
20	PLP4	Perceptual Linear Predictive4	PLP4 is the smoothed LPC cepstrum of PLP2.
21	TO	Temporal Energy Operator	Tempo, a time domain feature, which is simple to extract and have easy physical interpretation, like: the energy of signal and it helps in the study of tone and intonation.
22	TEO	Teager Energy Operator	Teager Energy Operator mainly shows the frequency and instantaneous changes of the signal amplitude that is very sensitive to subtle changes.
23	Kurto	Kurtosis	Kurtosis represents the "peakedness" of the LTAS,with steep distributions producing positive kurtosis values, and flat distributions producing negativevalues.
24	Skew	Skewness	Skewness describes the spectral tilt of the LTAS(long-term average spectrum) (negative or positive). Positively skewed distributions appear to have a "tail" of scores or values extended to the right of the bell curve. Negatively skewed distributions have a tail extending from the left of the bell curve.
25	Max Amplitude	Maximum Amplitude	Max Amplitude defines the maximum distance between the resting position and the maximumdisplacement of the wave.
26	Min Amplitude	Minimum Amplitude	Min Amplitude defines the minimum distance between the resting position and the minimum displacement of the wave.
27	5^{th} Moment	5^{th} Moment	5th moment about the mean for a sample i.e. array elements along the specified axis of the array.
28	6^{th} Moment	6^{th} Moment	6th moment about the mean for a sample i.e. arrayelements along the specified axis of the array.

For this paper we have prepared the data set by taking the average length of a conversation between the trained person and Computer via microphone using Praat software at 22500 Hz modulation. The average duration was 2min and 30s for all the 11 vowel utterances. The longest utterance duration was 2 min and the shortest utterance has taken of around 3s. The shortest utterance is not clear with clear voice, which may be for some technical difficulties up stretched in the sound tracking procedure.

B. Tools used during Data Collection

We have used some tools for collecting the speech data without which the whole procedure was impossible. The used tools are described in a table format as given below.

TABLE III. DESCRIPTION OF THE DATA COLLECTION USED IN THIS STUDY

Voice Recording Tools	Voice Recording Time	No of Recorded Speech	No of Speaker
Intel i3Processor	Avg time2 min 30 s	11 Assamese Vowels	5 persons
Praat Recording Software	Longest time 2 min	11 vowel utterances per each speaker	
Microphone with 22500 Hz modulation	Shortest time 3s	Total 55 utterances	

In general, telephonic or other stereo audio channels the maximum utterance duration is 4 min, the minimum1.32 min, and the mean 3.25s found in network traffics. As a humane being from the same native place we can easily understand the utterances but a person not familiar with the Assamese languages or a Machine can not recognize the dialectal of the individual utterer, so voiced features without any related data is the individual data that we have used in this research. As in Assamese language there are total 11 vowels are present therefore, all total we collected 55 utterances, where 11 utterances were from 5 person each. As we have collected total 55 utterances from 5 different trained person having 11 utterances per head and we are taking 28 vocal features for analysis of speech so our dataset consists of 55 rows and 29 columns (55 rows for the utterances, 28 columns for 28 features and one more column for storing the vowel samples). The dataset is given below.

TABLE III. CSV FILE OF ANALYTICAL VALUES OF 28 SPEECH FEATURES

Vowels_sa	mfcc	Mcc	fcc	MELBS	HFBS	LFBS	harmonicit	Modulatio	MSMC in j	fundament	mean freq	median fre	standard d	formant fr	formant fr	Zero crossi	pip1	pip2	pip3	pip4	temporal e	teager ene	kurtosis	skewness	Max	Min	5th moment	6th Moment
Vowel1_s1 (20, 152)	-5.26E-06	1.71E-06	(0.2057664	(0.0225256	(0.0078273	(77760,)	0.070785	-0.99101	12	4.07E-05	0.000175	2182.725	0.00754	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel1_s2 (20, 94)	-4.95E-06	-1.94E-05	(17.212846	(0.0051544	(0.0248802	(48096,)	0.070785	0.167269	24	0.005424	0.023112	2182.776	0.001724	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel1_s3 (20, 122)	-4.24E-06	-2.24E-06	(-26.81371	(0.0079052	(0.0109643	(62496,)	0.070785	0.989802	6	0.005673	0.024363	6365.389	0.000554	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel1_s4 (20, 134)	9.82E-06	-2.97E-05	(-32.46452	(0.0092920	(0.0041164	(68544,)	0.070785	0.334904	10	0.001638	0.007048	2182.736	0.003241	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel1_s5 (20, 68)	-2.05E-05	-6.12E-07	(-27.09494	(0.0101609	(0.0345427	(34560,)	0.070785	-0.91640	8	0.022288	0.099288	2182.826	0.001363	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel2_s1 (20, 91)	1.86E-05	-3.77E-06	(-20.05962	(0.0714859	(0.0010959	(46080,)	0.070785	0.942717	7	0.007192	0.03087	2182.781	0.009206	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel2_s2 (20, 102)	1.17E-05	-3.44E-06	(-37.70540	(0.0022637	(0.0067626	(51840,)	0.070785	-0.99101	710	0.000127	0.000545	2182.766	0.003375	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel2_s3 (20, 76)	1.40E-05	8.05E-07	(27.152695	(0.0601941	(0.0471228	(38880,)	0.070785	0.140609	6	0.00824	0.025323	2182.807	0.001186	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel2_s4 (20, 57)	-6.82E-06	-4.10E-06	(28.844168	(0.0526282	(0.0284717	(29088,)	0.070785	0.221635	5	0.00141	0.000048	2182.863	0.001222	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel2_s5 (20, 105)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel3_s1 (20, 91)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel3_s2 (20, 192)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel3_s3 (20, 70)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel3_s4 (20, 132)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel3_s5 (20, 68)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel4_s1 (20, 55)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel4_s2 (20, 137)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel4_s3 (20, 67)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel4_s4 (20, 75)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel4_s5 (20, 56)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel5_s1 (20, 49)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel5_s2 (20, 124)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel5_s3 (20, 114)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel5_s4 (20, 92)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel5_s5 (20, 83)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel6_s1 (20, 83)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel6_s2 (20, 89)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel6_s3 (20, 81)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel6_s4 (20, 150)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel6_s5 (20, 48)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel7_s1 (20, 75)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel7_s2 (20, 84)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel7_s3 (20, 87)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel7_s4 (20, 104)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel7_s5 (20, 59)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel8_s1 (20, 77)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel8_s2 (20, 100)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel8_s3 (20, 53)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel8_s4 (20, 74)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel8_s5 (20, 55)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel9_s1 (20, 46)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel9_s2 (20, 79)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel9_s3 (20, 56)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel9_s4 (20, 92)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel9_s5 (20, 55)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel10_s (20, 75)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel10_s (20, 71)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel10_s (20, 64)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel10_s (20, 71)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel10_s (20, 44)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel11_s (20, 67)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel11_s (20, 98)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel11_s (20, 49)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel11_s (20, 75)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vowel11_s (20, 47)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

IV. The Proposed Vowel Recognition Model

The proposed spoken vowel recognition model can be described by the block diagram in Figure1.

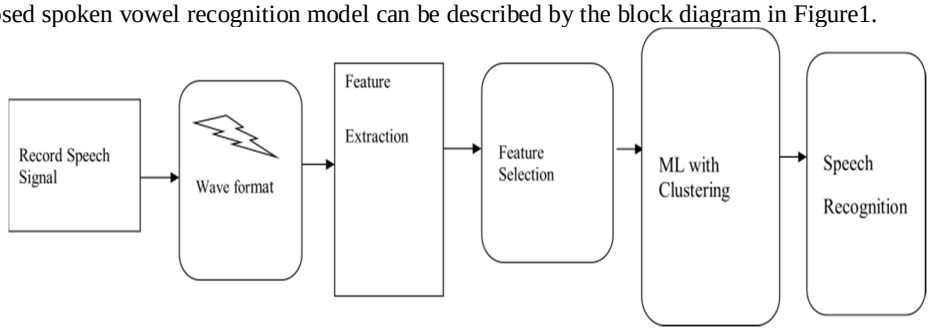

Fig.1.Architecture of Speech Signal Processing over the Network

A. Unsupervised Machine Learning Module

To achieve proficient clustering outcomes for unlabelled data Kmeans clustering is mostly used to cluster as it's simple, implementation friendly and the probability of getting the required cluster amount. Also, Kmeans clustering can be implemented on multivariate time series having each time point as a vector.In such case, to operate the k-Means clustering algorithm to cluster guests based on the time they spend at attractions for which the cluster labels are used as symbols to encode the time series of instance. It helps in some practical areas like, to cluster people with in a particular group where lots of network traffic occurs. After finalizing our dataset we used Clustering technique for modeling our system. Kmeans, a clustering technique that is famous for unlabelled data has been chosen by us. K-Means is a non-hierarchical data clustering method that falls under the category of centroid-based clustering [6]. Moreover, k-Means is used in a comprehensive outward record to split all the positions into a quantity of sets with unique variation rates for each [11]. The clusters we have found after programming in python are as below in Figure2.The scattered diagram obtained after clustering is given inFigure3.

```
array([[12.12991977,  8.1410154 ],
       [ 1.08849696,  1.05184601],
       [ 6.55680461,  6.67404195],
       [ 9.79322041, 10.05888546],
       [12.75515374,  5.618632  ],
       [ 5.09868155,  5.07492087],
       [ 7.66462029,  1.37168431],
       [10.02541923,  5.7298376 ],
       [11.47321598,  0.95636506],
       [11.13462606,  4.16166395],
       [ 9.032562  ,  3.35476144]])
```

Fig.2 Vowel Clusters

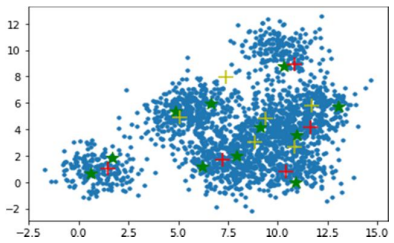

Fig.3. Scattered Chart of the Clusters

V. Result Analysis

In our designed system result is the main factor depending on which, we can further develop our system and the result we found are nothing but clustering accuracy. We have computed the accuracy with the increasing number of features, starting with the first five features then adding next five features serially up to the last feature. The accuracies we have found from our selected features are listed in the Table IV below.

Table IV: Accuracy Comparison of Features

Clustering with selected Features	First Five features	First Ten Features	First Fifteen Features	First Twenty Features	First Twenty Five Features	All Twenty Eight Features
Accuracies of Clustering	85	82	77	76	74	62

VI. Conclusion

We can explore conclusion based on the table mentioned above in the result and analysis part. In the table we have found different accuracies for different feature scales. The main conclusion Observed is that with increasing no of features accuracy of clustering has been decreasing and vice versa. Another conclusion that can be drawn is that scaling and scheduling of the features matter the accuracy values as some statistical features along with basic speech features have been used in our system. The statistical features are not so effective to achieve better accuracy in our system.

Future Work

In the research area of speech, the ASR has been developing from the early handy microphones to far-field microphones day by day. The increasing demand from users to make interaction with devices starved of a handy microphone or wearing it. In modern days now, across the world some famous apps are such as, Amazon's Echo, Google Home, Echo dot, Alexa, Amazon Alexa etc., have been used by many people at home. But some obscurities concealed in such handy situations also exist outside while using far-field microphones. Because, when the voice has reached the microphone, in the situation of far-field, the speech signal energy subsequently decreased. Relatively, the intrusive signals like echo, background noise, and dialogs from other utterer are so clear sometimes that we have no option to ignore. Some uncovered queries of ASR systems are like: (1) Can we develop additional information, such as language model by feeding information from the decoder back to the speech enhancement and separation component, if yes how it's possible? (2) Another, is the continuous optimization policy preferred with its mentioned easiness and optimal combination features, in case we require only optimization for the interpreted outcomes with adequate trained information. (3) Lastly, is it possible to design a model without robustness for us to propose? Moreover, ASR applications including syllabification using many standard speech coding algorithms for better noise free speech recognition is our future scope.

References

[1] Ljubomir Josifovski et al., "Robust Automatic Speech Recognition with Missing and Unreliable Data" in August 2002.

[2] Saliha Benkerzaz, Youssef Elmir, Abdeslam Dennai, "A Study on Automatic Speech Recognition", DOI:10.6025/jitr (Journal of Information Technology Review)Vol10,3August2019,P77-85.

[3] Dong Yu and Jinyu Li, "Recent Progresses in Deep Learning Based Acoustic Models" by IEEE/CAAJ. Autom.Sinica,VOL.4,NO.3,JULY2017,P396-409.

[4] Hicham Atassi et al. "Emotion Recognition from Spontaneous Slavic Speech", in Cog Info Com 2012,3rd IEEE International Conferenceon Cognitive Info communications, December25,2012P389-394 .

[5] MARK W. JONES and XIANGHUA XIE, "Clustering and Classification for Time Series Data in Visual Analytics: A Survey", published in IEEE ACCESS multi-disciplinary open access journal on December 10,2019, Digital Object Identifier10.1109/ACCESS.2019.2958551,vol7,pno:181314-181338.

[6] Frank Wessel and Hermann Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition", IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 1, January 2005.

[7] https://en.wikipedia.org/wiki/Assamese_language. Accessed online10th January2020

[8] 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

[9] Y.Hu and P.Loizou, "Evaluation of objective quality measures for speech enhancement", IEEE Transactions on Audio, Speech and Language Processing, Vol.16,No.1, pp.229-238,Jan.2008.

[10] Ganchev, I. Potamitis, N. Fakotakis and G. Kokkinakis, "Text-Independent Speaker Verification for Real Fast-Varying Noisy Environments", International Journal of Speech Technology, Vol.7,No.4,pp.281-292,Oct.2004.

[11] 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 Identifier10.1109/ACCESS.2019.2958551,volume7,pno:181314-181338.

[12] https://en.wikipedia.org/wiki/Assamese_language

[13] Dibya Jyoti Bora and Anil Kumar Gupta, "A Comparative Study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm", International Journal of Computer Trends and Technology,Vol.10,No. 4,pp.108 113, 2014

[14] T.Vogt and E. Andre, "Comparing Feature Sets for Acted and Spontaneous Speech in view of Automatic Emotion Recognition", Proc. IEEE,2005