IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Comparison of Machine Learning Algorithms for Speech Recognition of Endangered Gadaba Language

A Study of the Signal Processing aided Feature Extraction based Speech Recognition

[1]S. Ramadevi, [2]C. Sushmita, [3]B. Reshma, [4]B. Himaghna, [5]G. Likhitha, [6]E. Jerina, [7]D. Kavitha Devi,

[8]C. K. Abhishek

[1]Assistant Professor, [2,8]Project Guides, [3,4,5,6,7]Students
[1,2,3,4,5,6,7]Electronics and Communication Engineering, Andhra University College of Engineering for Women(AUCEW), Visakhapatnam, India
[8]PhD in Department of Linguistics, Andhra University, Visakhapatnam, India

Abstract: We compare the various ML (Machine Learning) based speech recognition algorithms and for a Gadaba Creole Speech recognition and translation towards Language Learning to promote Gadaba Language Revival through technology and sophistication. We study the efficiencies of various feature extraction signal processing techniques such as zero crossing rate, pitch, MFCC, frequency shift for three Machine Learning based Statistical speech recognition Algorithms: HMM (Hidden Markov Model), GMM (Gaussian Mixture Model) Cross Correlation Routines.

We have trained the algorithms with 20 Gadaba words towards isolated Gadaba word and phrase recognition and eventual English Translation with POS (Parts of Speech) consideration. The Generalized Speech Recognition Algorithms allowed English Speech Recognition and its Gadaba Translation as well. The accuracies of the Algorithms are studied over various features extracted enabling audio feature selection for Gadaba Language.

Index Terms **- Feature Extraction, Signal Processing, Machine Learning (ML), HMM (Hidden Markov Model), GMM (Gaussian Mixture Model), Cross Correlation, Gadaba Language, Speech Recognition, Translation**

I. INTRODUCTION

Signal Processing techniques employed for feature extraction of speech signals find applications specific to the audio features extracted. For example, Isolated word recognition, speech segmentation [1], word boundary identification [2], denoising, stop consonant identification [3] can be achieved by extracting time domain features such as zero crossing rate (ZCR) and short term energy along with frequency-domain features, such as spectral centroid and spectral spread. While for biometric speaker identification, Mel frequency cepstral coefficients (MFCC) and ZCR features [4] form the speech template for the system algorithm. Speech emotion recognition [5] need a cluster of acoustic features such as lower order MFCC, Discrete Wavelet Transform (DWT), Linear Prediction Cepstral Coefficient [6], Cross correlation [7] for SVM (Support vector machines) classifier, pitch, energy and ZCR to be extracted for better accuracy with extensive speech information. Acoustic features such as pitch, HPS (Harmonic Product Spectrum) and MFCC aid in Gender identification applications [8]. For speech recognition in non-stationary disturbances (non-white noise) such as background music or background speech, acoustic power features such as PNCC [9](Power Normalized Cepstral Coefficient) and power flooring techniques could be employed.

Based on the feature extraction various acoustic modeling algorithms can be chosen such as K-nearest neighbors (KNN)[5], Hidden Markov Model (HMM)[10], Convolutional Neural Networks for continuous speech recognition[11], Support Vector Machine (SVM), Artificial Neural Networks (ANN), Autocorrelation, Cross Correlation and Gaussian Mixtures Model (GMM), Spectrum Normalization, Wiener Filter.
Further, a language model either statistical (n-gram) or neural network based is used to convert speech data into a more meaningful sequence POS (Parts of Speech) scoring based on a corpus of text from the language of choice.
The efficiency of Statistical or AI(Artificial Intelligence) based Machine Learning Algorithms rely on the features on which the training and testing (speech template matching) is performed.

Gadaba is a spoken tribal language of the Eastern Ghats of India. Today, with development and communication prioritizing majority Languages, the inhabitants of the regions have alienated the language and shifted towards the dominant language spoken in their study/workplace. With less than 1000 speakers inclusive of various Gadaba language creole with no written form, the future of Gadaba is even more compromised. Further, this low resource language lacks the large amount of transcribed speech needed to train algorithms.

In this paper, we compare the most prominent Statistical Speech Recognition Algorithms in Literature namely Hidden Markov Model, Gaussian Mixture Model, Minimum Error Model by training and testing the Gadaba audio data through audio feature extraction of pitch, zero-crossing rate, turns count, RMS vale, MFCC (Mel Frequency Cepstral Coefficients), Cross Correlation.

II. SPEECH RECOGNITION ALGORITHMS AND FEATURE EXTRACTION TECHNIQUES

2.1 Speech Recognition-Mathematical Formulation

The statistical formulation of speech recognition can be understood as the most likely sequence of words $w = w_1, w_2, \dots w_n$ given a set of time/frequency domain acoustic features extracted as $F = F_1, F_2, \dots F_L$, given in Eq. 2.1, Eq. 2.2

$$\widehat{w} = \arg\max_{w} \{P(w/F) \tag{2.1}$$

$$\widehat{w} = \arg\max_{w} \{P(F/w)P(w)\} \tag{2.2}$$

The prior probability $P(w)$ of word occurrences can be evaluated using an efficient language model (such as n-gram model) and training data. While an efficient acoustic modelling with reliable feature extraction can enable the computation of $P(F/w)$

2.2 Time & Frequency Domain Features: RMS, Zero-Crossing Rate, Turns Count, Cross Correlation

The Speech signal waveform can be analyzed using the root mean square value computed dynamically over M samples of equal weightage of a defined causal window is given by Eq. 2.3

$$RMS(n) = \left(\frac{1}{M}\sum_{k=0}^{M-1} x^2(n-k)\right)^{1/2} \quad M \ll N \tag{2.3}$$

Zero crossing rate is the frequency with which the signal transitions between the regions separated by the zero state (activity) line. Computed over a moving window, this parameter is affected by base-line drift, DC-bias and any low frequency signal artifacts. High pass filtering operation like finding the derivative of the speech signal can remove the affect of DC bias. The bandwidth of the speech signal is fixed with particular band of Zero-Crossing Frequency, while noise is different frequency spectra, which can help in speech versus silence decision.

Depending on the source of the signal/starting point, voiced and unvoiced signals can be differentiated based on their intensity (RMS) as well as the frequency content (Zero-Crossing Rate and Turns Count). Voiced speech is obtained by modulating the limited bandwidth wave from the glottis by the transfer function of the vocal tract. While for unvoiced speech the turbulent wind which is of white spectra (constituting all the frequencies) coming from the lungs is forced through the vocal tract, whereby the spectra of the vocal tract is obtained. Therefore, compared to the unvoiced sound spectrum, the voiced spectrum contains limited and low frequency components thus the zero-crossing rate is low for voiced speech.

Turns count is used measure the activity of the speech signal, here we look at the local extremum for determining the number of turns or phase change and gradient change of the signal with respect to time. In the interference pattern with a moving window, the occurrence of the number of peaks is the turns count. The prominent turns greater than a threshold vale are selected to make the parameter resistant to noise fluctuations.

Voiced Part of the signal is identified as the portion of the sound waveform for which RMS is greater than some threshold decided by the noise and signal amplitude and the Zero Crossing Rate is less than some threshold decided by noise and signal frequencies. The Unvoiced part of the sound waveform is identified as the part of the audio signal for which there is an appropriate low RMS and high Zero Crossing rate and turns count.

The cross correlation between two sequences x, y is given by Eq. 2.4, Eq. 2.5

$$\gamma_{yx}(l) = \sum_{n=-\infty}^{\infty} x(n)y(n-l), l = 0, \pm1, \pm2, \pm3 \dots \tag{2.4}$$

$$\gamma_{yx}(l) = x(l) * y(-l) \tag{2.5}$$

For two jointly stationary random process, x_n, y_m , the cross correlation is the expectation of product of x_n and complex conjugate of y_{n-m}. The error against cross correlation symmetry is a useful feature for speech recognition.

2.4 Quefrency Domain Feature: MFCC

Frequency Spectrum of sound is a product of the frequency spectra of the glottal pulse wave and the vocal tract frequency response. Glottal pulse, a high frequency signal generated by the vocal folds responsible for pitch of the sound produced, the frequency

response of the vocal tract which acts as a filter caries the information of the timber (actual) of the sound. The Frequency spectrogram is transformed into perceptually relevant Mel Scale (as humans perceive frequency logarithmically), representing the characteristics of the sound phoneme given by Eq. 2.6

$$f_{mel} = 2595 log_{10}\left(1 + \frac{f}{700}\right) \tag{2.6}$$

Cepstrum (Eq. 2.7) is obtained by applying inverse fourier transform to log amplitude spectrum. Discrete cosine transform is applied instead of inverse Fourier transform since the former give real valued coefficients.

$$c(n) = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x(n)\}|\} \tag{2.7}$$

MFCC coefficients (Eq. 2.8) parametrically map cosine waves with a particular frequencies to the log spectrum. Frequency values of the sound spectrum on the higher end contain information of the glottal pulse. The first 13 coefficients on the lower end of the frequency spectrum along with their first and second derivatives provide the information of the spectral envelope/formants/phonemes, which is of interest for ML based Speech Recognition. For obtaining MFCC, we convert the log spectrum to Mel Scale before applying the Discrete Cosine Transform

$$MFCC(m) = \frac{1}{R}\sum_{r=1}^{R} \log\left(MF_{\dot{n}}(r)\right) \cos\left(\frac{2\pi}{R}\left(r + \frac{1}{2}\right)m\right) \tag{2.8}$$

$MF_{\dot{n}}(r)$ is the frequency spectrum converted to Mel Scale, R is the Mel Filter Count. Fig. 1 shows the Mel Scale Triangular Filter Banks

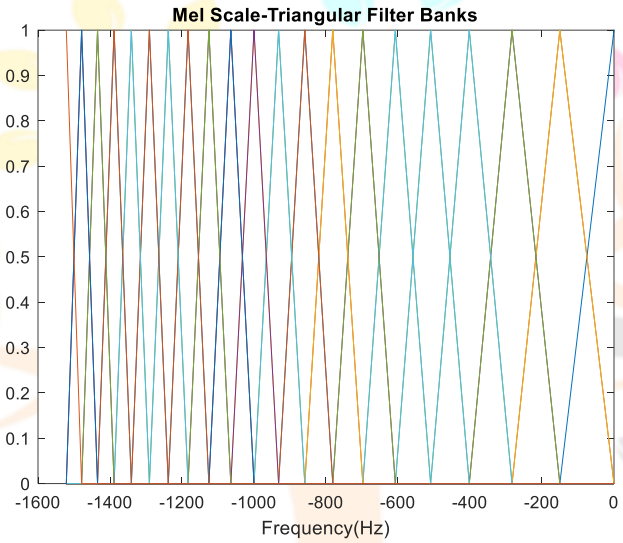

Figure 1. Mel Scale Triangular Filter Banks

2.3 Hidden Markov Model (HMM) Algorithm for Speech Recognition

For a first order Markov Process/Markov Chain, the probability of the present state depends only on the immediate previous state given by Eq. 2.9

$$P(X_n = x|\, X_1 = x_1, X_2 = x_2, \ldots X_{n-1} = x_{n-1}) = P(X_n = x|\, X_{n-1} = x_{n-1}) \tag{2.9}$$

But in Speech Recognition systems all states are not observable directly, they are hidden. For this case, we employ Hidden Markov Model (Fig. 2) in Speech Recognition. For example, the parts of speech (POS) of the uttered sentence is not directly observable. Using the emission probability $P(X_i/S_i)$which is the probability of the given observation X_i given a state S_i and the transition probability $P(S_i/S_{i-1})$ which is likelihood of change of one state S_{i-1} to another state S_i, the probability of the observations is given by Eq. 2.10, Eq. 2.11

$$p(X) = \sum_s p(X/S)\, p(S) \tag{2.10}$$

$$P(X/S) = \prod P(X_i/S_i),\ P(S) = \prod P(S_i/S_{i-1}), \tag{2.11}$$

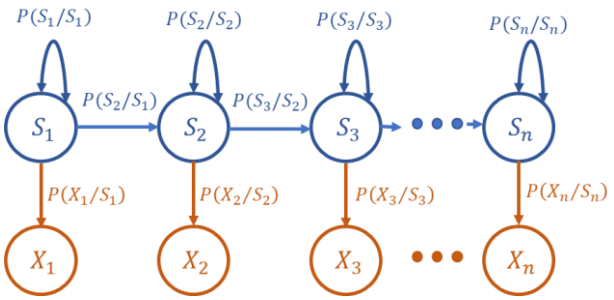

Figure 2. Hidden Markov Model represented with State transition Probabilities and Observation to State Emission Probability

The Hidden/internal states *VT* can be decoded using the Viterbi Algorithm by maximizing the joint probability of the observations and most likely state sequence given in Eq. 2.12

$$VT(j) = \max_{S_0, S_{0,}, \ldots S_{t-1,}} P(S_{0,}, S_{1,\ldots} S_{t-1,} x_{1,} x_{2,} \ldots x_{t,}, S_t = s_j | \alpha) \tag{2.12}$$

2.4 Gaussian Mixture Model (GMM) Algorithm for Speech Recognition

Gaussian Mixture Model (Fig. 3) is a multimodal-multivariate Gaussian Distribution which can be used to model the 39 features of MFCC using 39 variables. The probability of the observed feature vector modelled by GMM is given by Eq. 2.13

$$p(x|S = i) = \sum_{n=1}^{M} c_{in} \mathcal{N}(x: \mu_{in} \Sigma_{in}) \tag{2.13}$$

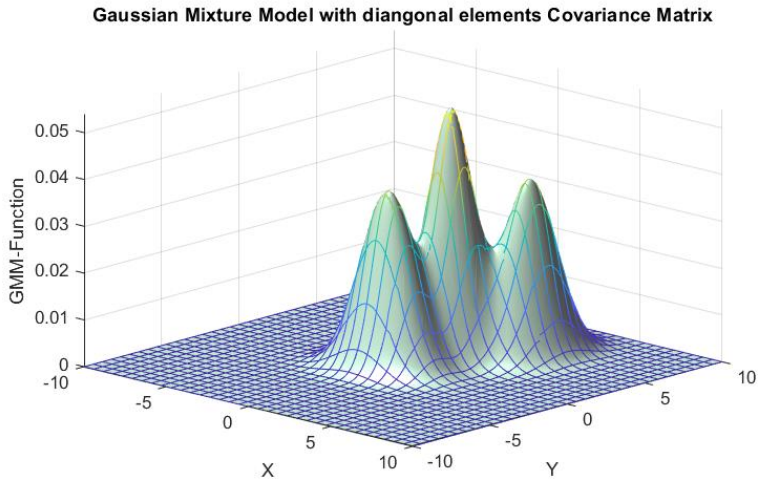

Figure 3. Gaussian Mixture Distribution for $\mu = [1\ 2; 0\ -1;\ 3,5]$, $\sigma(1,:,1) == [1,1]$, $\sigma(1,:,2) == [2,1]$, $\sigma(1,:,1) == [1,2]$, and mixture proportion p=[1,1,1];

c_{in}, μ_{in}, Σ_{in} are coefficient (mixing parameter-deciding how of each component), mean and covariance matrix of the Gaussian distribution n respectively.
Akaike information criterion (AIC) is given by Eq. 2.14

$$AIC = (2 * NlogL) + (2 * p) \tag{2.14}$$

where, $NlogL$ is the negative log likelihood, p is the number of parameters.
Bayes information criterion (BIC) is given by Eq. 2.15

$$BIC = (2 * NlogL) + (p * \log(n)) \tag{2.15}$$

These criteria are used for finding how well the given data can fit to the GMM model. A lower AIC and BIC account to a better fit. The Mahalanobis distance measures the distance between the trained GMM model and test speech vector y, in terms of the number of standard deviations the data is from the mean of the GMM distribution. The Mahalanobis distance is given by Eq. 2.16

$$d_M = \sqrt{(y_{MFCC} - \mu_{GMM}) \sum^{-1} (y_{MFCC} - \mu_{GMM})'} \tag{2.16}$$

where, y_{MFCC} is the test speech MFCC feature vector, while μ_{GMM} is the mean of the GMM model trained.

III. RESULTS AND TRAINING AND TESTING PROCEDURE FOR SPEECH RECOGNITION

3.1 The Gadaba Language Training Procedure implemented in MATLAB

For the HMM Algorithm, the sampling frequency, recording time, number of Gadaba words, number of samples for training are initialized. The reference word is recorded and is converted into frequency spectrum. The real part of the Word Spectrum is considered and normalized. Around 20 Gadaba words are recorded 3 times each (based on the number of samples initialized) and their respective normalized real part of the frequency spectrum are obtained. The error against cross correlation symmetry of the training data with the reference word is evaluated, this is used as the audio feature for Gadaba word recognition. The mean of the audio feature for each Gadaba word is found by averaging over the results of the number of samples for each word.

An equiprobable transition matrix is considered, while the emission probability matrix is constructed using normal distribution with the audio feature extracted employed as the mean of the distribution. The Training Model along with transition matrix, emission matrix and the normalized real frequency spectrum of the reference word are saved.

The HMM algorithm's Emission Probability matrix can also be constructed using various other audio features such as pitch (Fig.5), zero crossing rate and turns count for voiced and unvoiced distinction (Fig. 4) and also MFCC. Fig. 4 shows the various audio features extracted for the Gadaba words.

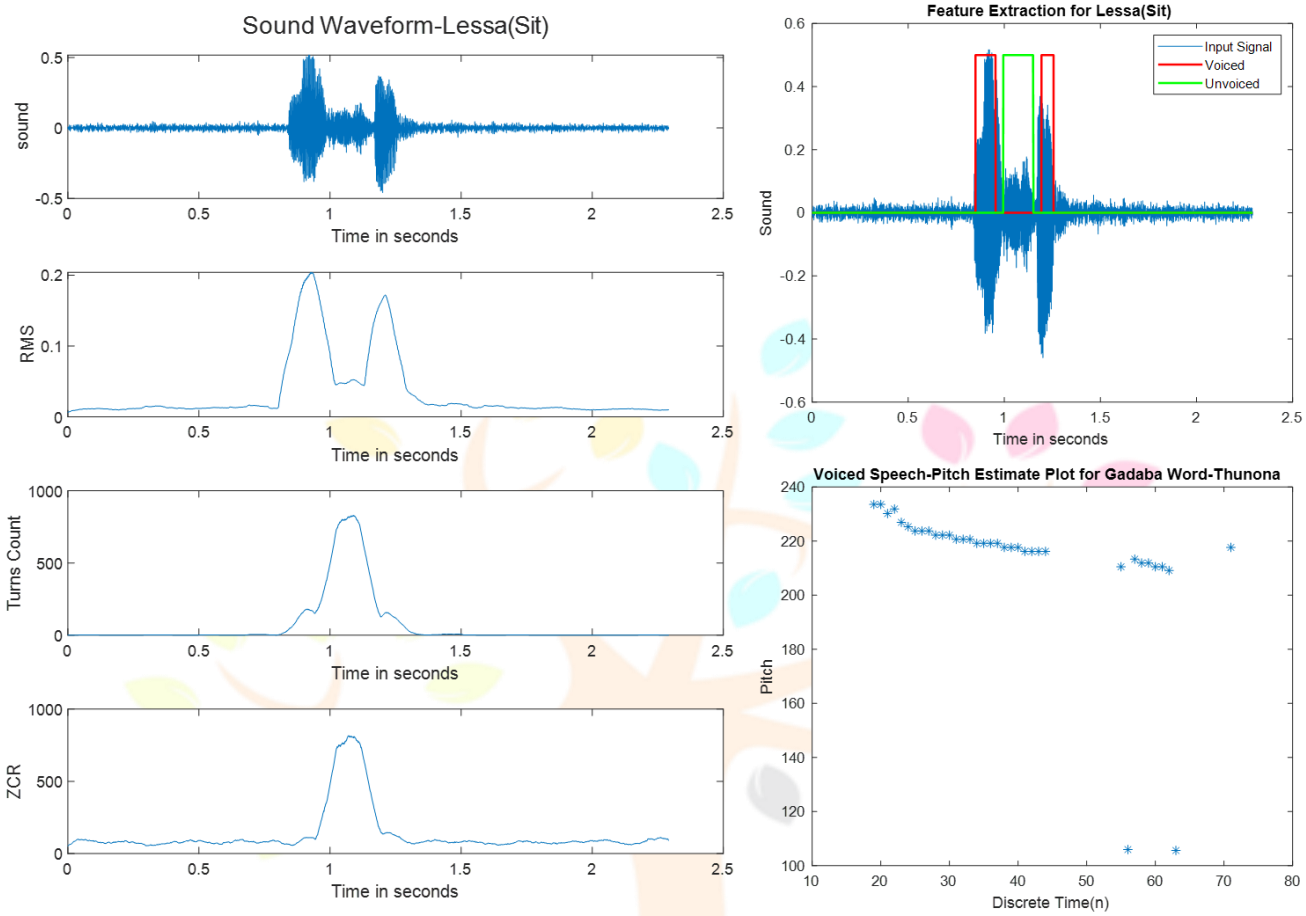

Figure 4. (Left) Sound Waveform, RMS, Turns Count, Zero Crossing Rate for Gadaba Word (Lessa). (Right-Bottom) Pitch for voiced speech by removing unvoiced speech components. (Right-Top) Segmentation of Sound Waveform into Voiced and unvoiced portions

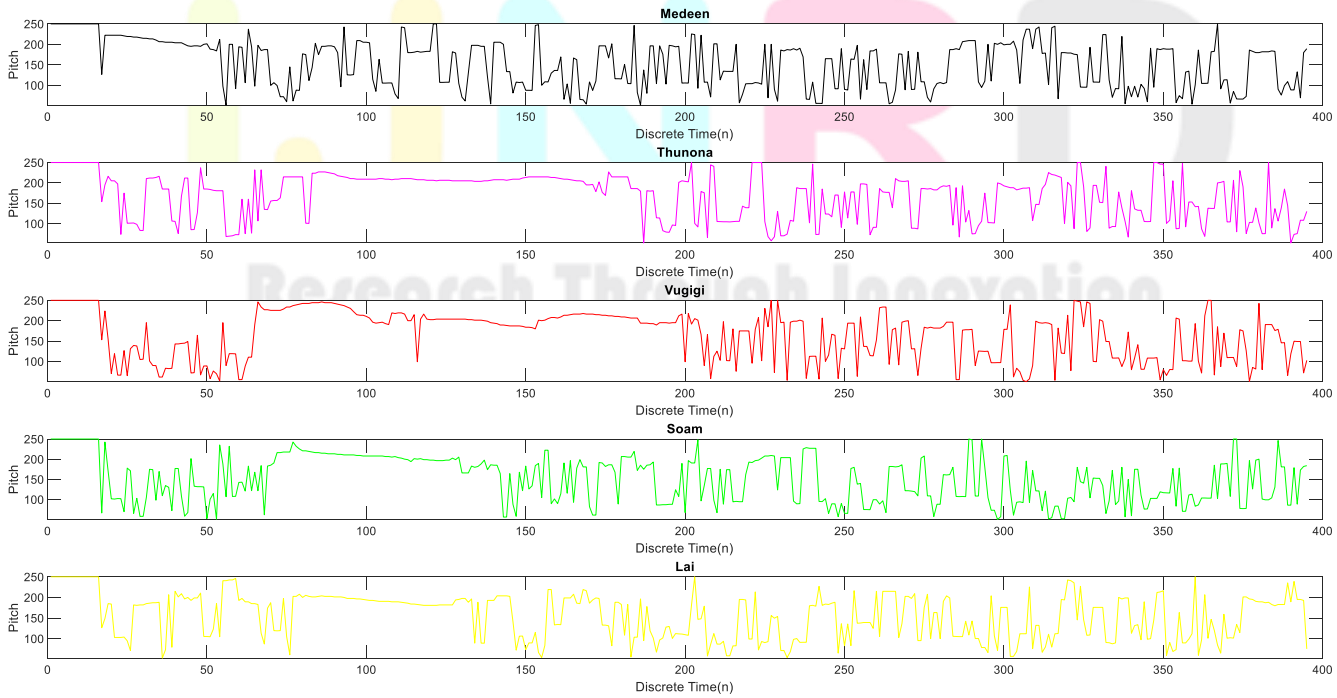

Figure 5. Pitch Waveforms for Different Gadaba Words-Medeen, Thunona, Vugigi, Soam, Lai

While for the Minimum Error against Cross Correlation Algorithm, the normalized real frequency spectrum of all the trained Gadaba words (Fig.6) are saved as the training Model to be retrieved during Voice Pass.

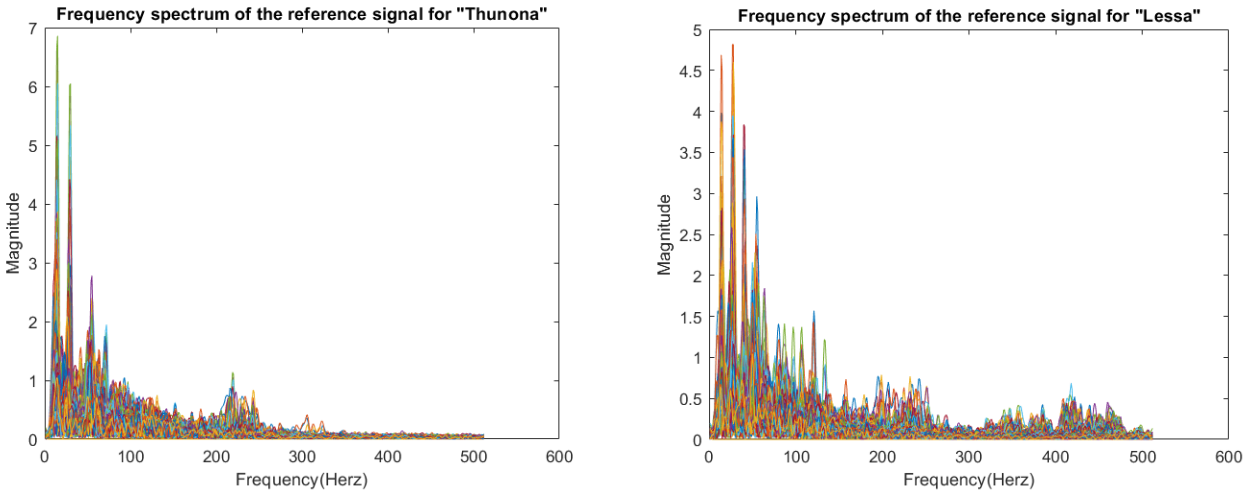

Figure 6. Frequency Spectrum of Spoken Gadaba Words -Thunona, Lessa

For GMM Algorithm, the MFCC features (Fig. 7) of the recorded Gadaba words are extracted Gaussian Mixture Distribution is constructed for each Gadaba Word by fitting the MFCC data to GMM model (Table 3.1). The GMM distribution for each Gadaba Word is saved as the training Model.

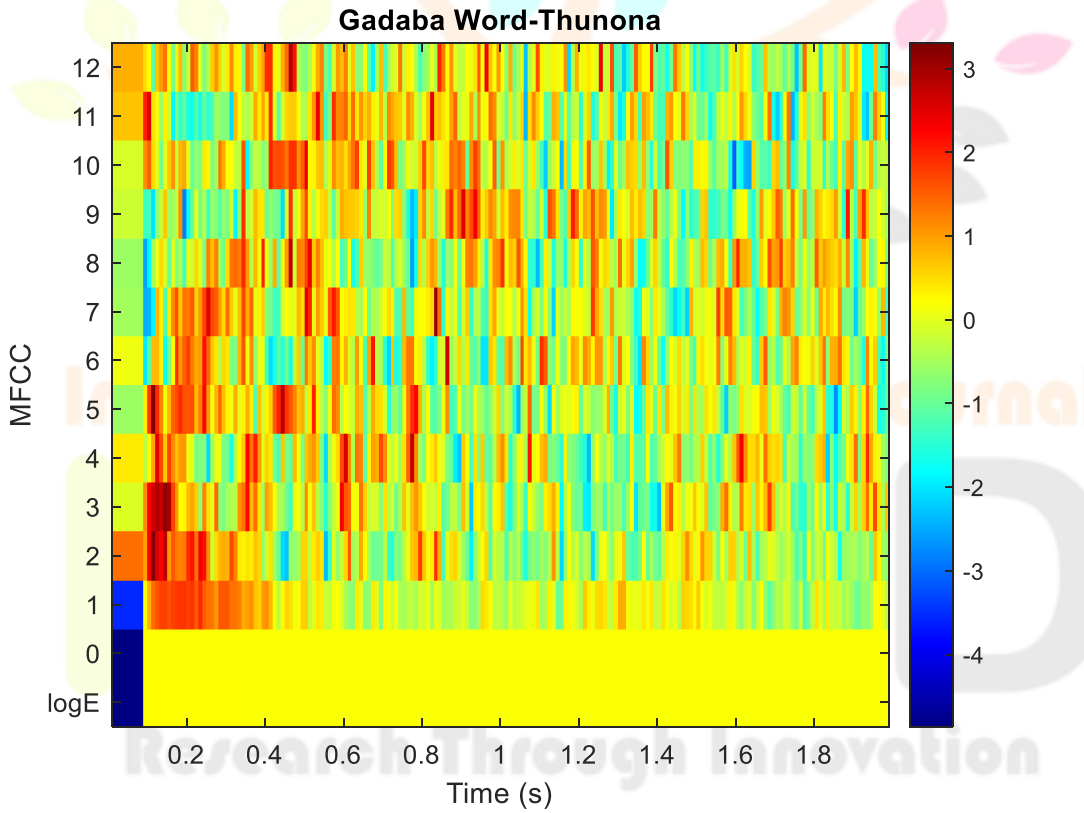

Figure 7. MFCC Spectrum for Gadaba Word-Thunona

Table 3.1: Gaussian Mixture Distribution constructed by fitting the MFCC data to the GMM model

Gadaba Words	Medeen	Thunona	Vugigi	soam	Lai	Tonta Bullu	Noomu	Bontelu	Gusso	Girgi datha
Number of Variables,	14	14	14	14	14	14	14	14	14	14
Distribution Name	GMM	GMM	GMM	GMM	GMM	GMM	GMM	GMM	GMM	GMM

Number of Components	1	1	1	1	1	1	1	1	1	1
Component Proportion	1	1	1	1	1	1	1	1	1	1
Shared Covariance	0	0	0	0	0	0	0	0	0	0
Number of Iterations	3	3	3	3	3	3	3	3	3	3
Regularization Value	0	0	0	0	0	0	0	0	0	0
Negative Log Likelihood	9.74E+02	8.53E+02	8.96E+02	9.11E+02	7.82E+02	9.48E+02	8.41E+02	8.06E+02	9.79E+02	1.06E+03
Covariance Type	'full'	'full'	'full'	'full'	'full'	'full'	'full'	'full'	'full'	'full'
μ	1x14 double	1x14 double	1x14 double	1x14 double	1x14 double	1x14 double	1x14 double	1x14 double	1x14 double	1x14 double
σ	14x14 double	14x14 double	14x14 double	14x14 double	14x14 double	14x14 double	14x14 double	14x14 double	14x14 double	14x14 double
AIC	2186.3	1943.0	2029.9	2059.6	1801.9	2133.9	1919.5	1850.7	2195.6	2350.85
BIC	2577.6	2334.3	2421.2	2450.9	2193.2	2525.2	2310.8	2242.0	2586.9	2742.15
Converged	1	1	1	1	1	1	1	1	1	1
Probability Tolerance	1.0E-08	1.0E-08	1.0E-08	1.0E-08	1.0E-08	1.0E-08	1.0E-08	1.0E-08	1.0E-08	1.0E-08

3.2 The Gadaba Word/Phrase Recognition Procedure-Voice Pass for test audio data (in MATLAB)

For the Gadaba word recognition using HMM Algorithm, the saved HMM model is loaded for emission and transition matrix data. The test audio data is recorded and converted into normalized real frequency spectrum. The error against cross correlation symmetry between the test audio and reference word is evaluated. This audio feature is inputted into the HMM trained model, and high probability Gadaba word is selected from the probability vector at the audio feature index as shown in the Fig. 8. The Excel sheet having the trained Gadaba words in order with English meaning are read according to the output recognized to display the recognized Gadaba word with English meaning.

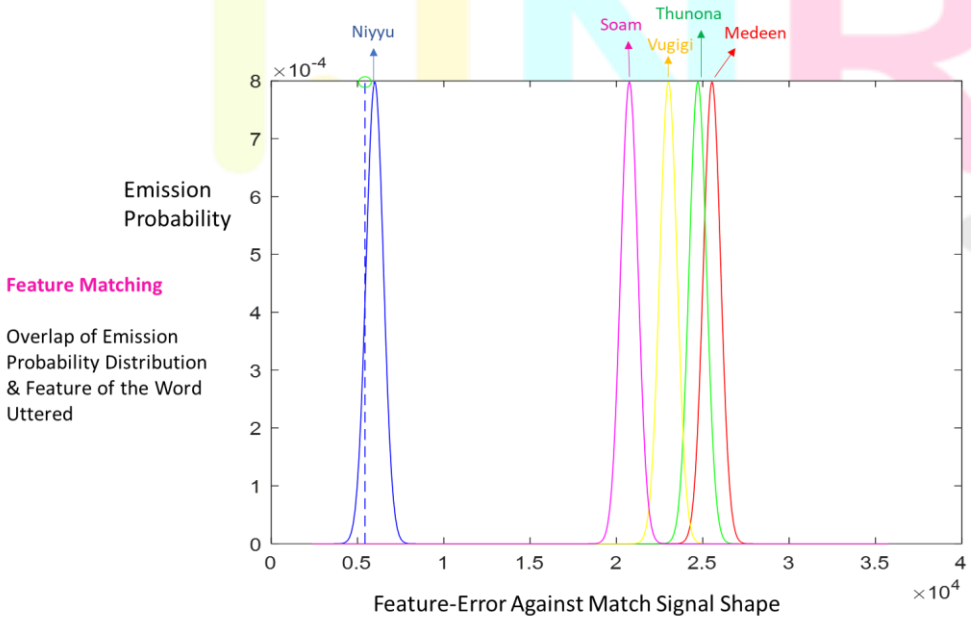

Figure 8. HMM Speech Recognition-Emission Probability Distribution overlapped with audio feature for Gadaba Word 'Niyyu'

For the Minimum error Algorithm, the error against the cross correlation symmetry for the test audio data with the normalized frequency spectrum of the trained words loaded are evaluated. The index of the trained Gadaba words with Minimum Cross correlation error with the test audio is found as shown in Fig.9. This index is matched to the Excel Training Word Sheet for Gadaba word/phrase recognition.

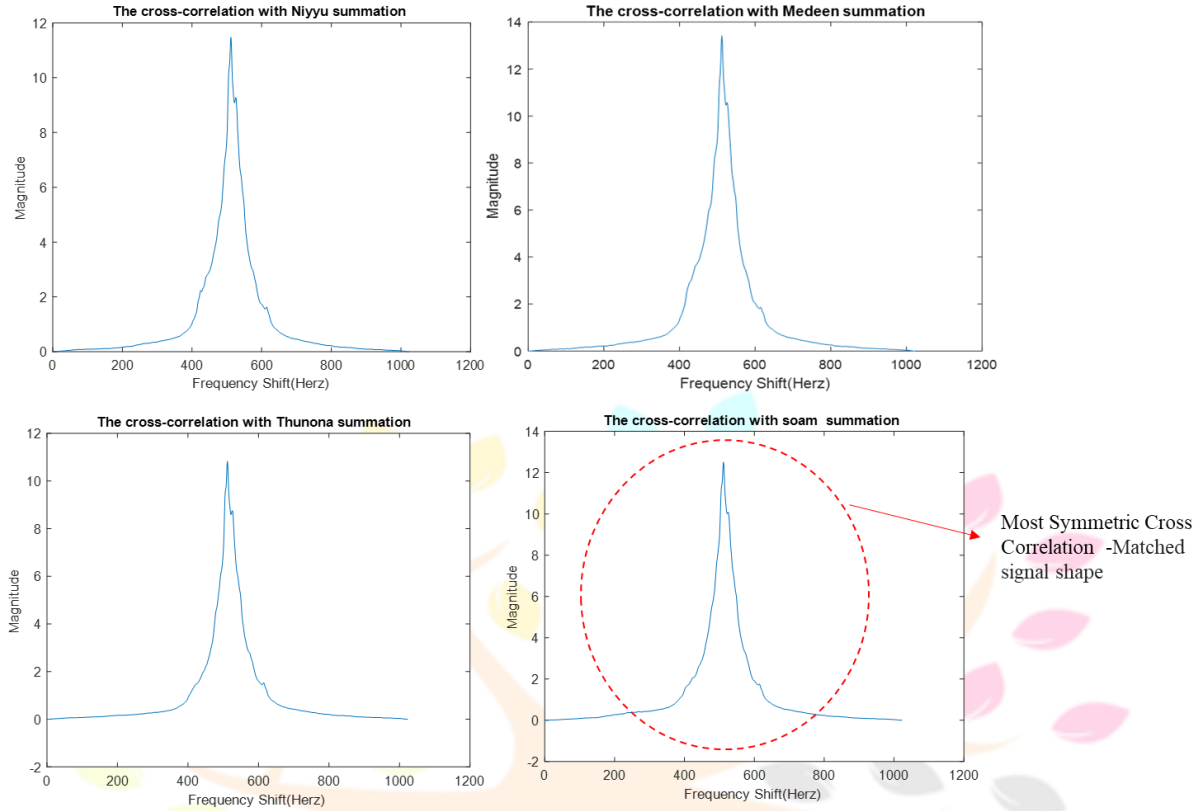

Figure 9. Minimum Error Algorithm based Speech Recognition-The Cross Correlation graphs of the test audio signal with the trained Gadaba words' frequency spectrum. The bottom right Cross Correlation with Gadaba word 'soam' is most symmetric with minimum error against symmetry.

For the GMM Algorithm, the MFCC of the test audio signal are evaluated and the minimum Mahalanobis distance (Table 3.2) to the Gaussion Mixture Model loaded for among the Gadaba words trained is evaluated and match to the Excel Training Sheet for Gadaba word/phrase recognition.

Table 3.2: Mahalanobis distance of the trained GMM distributions (for different Gadaba words) to the audio test data. The number of transitions between voiced and unvoiced speech for various Gadaba words

Gadaba Word	Mahalanobis distance $d_M = \sqrt{(y_{MFCC} - \mu_{GMM})\sum^{-1}(y_{MFCC} - \mu_{GMM})'}$	Transitions between voiced and unvoiced phonemes for each Gadaba Word
Medeen	1342.43	6
Thunona	1638.015	8
Vugigi	1325.211	10
soam	1149.299	4
Lai	1217.095	2
Tonta Bullu	1339.396	14
Noomu	1032.355	5
Bontelu	1439.116	7
Gusso	1185.37	3
Girgidatha	1410.017	16

The Gadaba Word and Phrase recognition and English Translation outputs at the command window with Parts of Speech Consideration are shown in Fig. 10, 11. Fig. 12 shows English speech recognition and translation to Gadaba Language. Table 3.3 show the accuracy of the various algorithms in noisy and noise free environments. The Minimum Error against Cross Correlation Symmetry Algorithm exhibited higher accuracy compared to the other algorithms implemented.

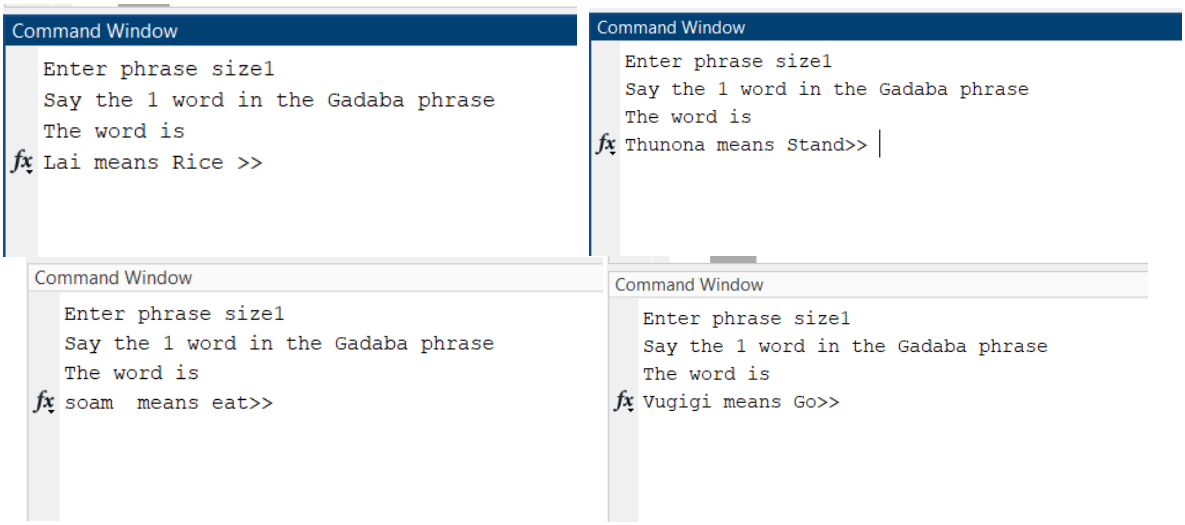

Figure 10. Command Window output of Gadaba Word Recognition and English Translation

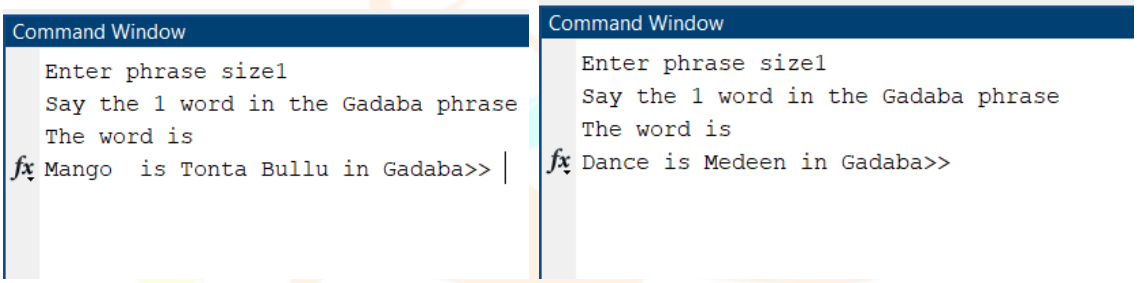

Figure 11. Command Window output of English Word Recognition and Gadaba Translation

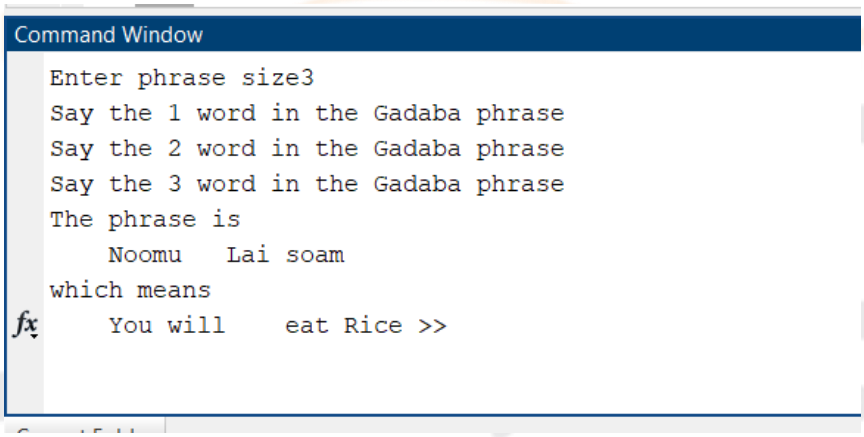

Figure 12. Command Window output of Gadaba Phrase Recognition with SOV (Subject-Object-Verb) in Gadaba to SVO (Subject-Verb-Object) English Translation.

Table 3.3: The Accuracy of Various Algorithms under noisy and noise less environment.

Accuracy in Gadaba Speech Recognition	HMM Algorithm	Minimum Error against Cross Correlation Symmetry Algorithm	GMM Algorithm
In noise free environment	60%	75%	65%
In the presence of white noise	50%	60%	53%

IV. ACKNOWLEDGMENT

We would like to sincerely acknowledge the valuable support and guidance of Professor S.K. Bhatti Madam, Principal, Andhra University College of Engineering for Women and the Head of the Department of Electronics and Communication Engineering, Andhra University College of Engineering for Women, Dr. S. Aruna Madam.

REFERENCES

[1] M. M. Rahman and M. A.-A. Bhuiyan, "Continuous bangla speech segmentation using short-term speech features extraction approaches," International Journal of Advanced Computer Science and Applications, vol. 3, no. 11, 2012, [Online]. Available: https://www.academia.edu/download/59515251/Cashless_Society_pg_197-20320190604-52015-1ydu74l.pdf#page=143

[2] J. Sangeetha and S. Jothi Lakshmi, "Robust automatic continuous speech segmentation for indian languages to improve speech to speech translation," Int. J. Comput. Appl. Technol., vol. 53, no. 15, 2012, [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=243c4da080a61bec6eb101520bb4a13f35bdc1e2

[3] N. N. Lokhande, N. S. Nehe, and P. S. Vikhe, "Voice activity detection algorithm for speech recognition applications," in IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012), vol. iccia, 2012, vol. 6, pp. 1–4.

[4] N. Chauhan, T. Isshiki, and D. Li, "Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database," in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Feb. 2019, pp. 130–133.

[5] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," Int. J. Speech Technol., vol. 23, no. 1, pp. 45–55, Mar. 2020.

[6] U. Jain, K. Nathani, N. Ruban, A. N. Joseph Raj, Z. Zhuang, and V. G.V. Mahesh, "Cubic SVM Classifier Based Feature Extraction and Emotion Detection from Speech Signals," in 2018 International Conference on Sensor Networks and Signal Processing (SNSP), Oct. 2018, pp. 386–391.

[7] S. Chandaka, A. Chatterjee, and S. Munshi, "Support vector machines employing cross-correlation for emotional speech recognition," Measurement, vol. 42, no. 4, pp. 611–618, May 2009.

[8] K. Patel and R. K. Prasad, "Speech recognition and verification using MFCC & VQ," Int. J. Emerg. Sci. Eng. (IJESE), vol. 1, no. 7, pp. 137–140, 2013.

[9] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 2010, pp. 4574–4577.

[10] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition," in Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005, 2005, pp. 799–804.

[11] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Convolutional Neural Networks-based continuous speech recognition using raw speech signal," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, pp. 4295–4299.