INTERNATIONAL JOURNAL OF FUZZY SYSTEMS and ADVANCED APPLICATIONS
DOI: 10.46300/91017.2020.7.1 Volume 7, 2020

Fuzzy Variable Frame analysis for Speech Recognition

VANI H Y , ANUSUYA M A
Sri Jayachamarajendra College of Engineering, JSS TI Campus,
Manasagangothri, Mysuru, India
vanihy@sjce.ac.in

Abstract**—**Recent works in machine learning has focused on models such as Support Vector Machine(SVM), Artificial Neural Network(ANN) and Long Short Term Memory (LSTM), for automatically controlling the generalization and parameterization of the optimization process. This paper presents a fuzzy interpretation frame analysis procedure using LSTM classifier for noisy speech at word level using thresholding and local maxima procedure at framing level for the recognition process. Front end MFCC procedure has been modified in the framing phase to reduce the number of noisy frames using thresholding at two level local maxima procedures. A comparative results of various classifiers like SVM with kernel function, ANN and LSTM are tabulated for recognition accuracies. A fuzzy interpretation at the framing level to calculate optimal frames has been presented in this paper. In the proposed work 20% of unwanted processing of frames is reduced that equally produces the accuracies obtained by fixed frame analysis. An investigation shows that the obtained features with LSTM decrease word error rate still by 1% as increasing the recognition accuracy from 98 to 99% . approach**.**

Keywords—Variable frame, frame selection, Threshold, Local maxima, fuzzy interpretation , Support Vector Machine(SVM), Artificial Neural Network (ANN) and Long short Term Memory (LSTM).

I. INTRODUCTION

SPEECH recognition [1] systems have become one of the leading applications for machine learning and pattern recognition technology. Speech recognition application task is difficult because of the variability that exists in speaking with every individual person, which is realized as a complex process. Along with this speech nature, nonlinearity dynamics also matters for the processing and analysis of speech signals. The speech signals are quasi-stationary and non-stationary in nature .To analyze these signals generally fixed frame and variable frame length procedures are used. Since vowels are stationary for long duration fixed frames are preferred and variable frame analysis is more suitable for consonants because of the existence of non-stationary property for longer duration. Hence these are proposed for speech recognition and speaker identification. In speech signal processing it is preferable to divide into frames to get stationary state. The procedure includes splitting of speech signals into frames and combining the frames into a speech signal. Several strategies exist in literature with fixed and variable frame length size. Few methods has been discussed in section 2 of literature survey. In the proposed work a fuzzy interpretation for selecting and processing the

optimal frames for noisy speech signal has been presented by adopting local maxima and threshold concepts with some of the machine learning classifiers like LSTM,SVM and ANN. The terms considered for fuzzy interpretation are low, medium and high that represents fuzzy frames towards the end, middle and to the beginning of the speech signal respectively. The rest of the paper sections are arranged as follows : Section 2 presents the literature survey in the area of variable frame selection. Frame selection procedure adopted in the proposed work is discussed in section 3.Section 4 describes about the data set used for experiments. Methodology and the experimental setup has been discussed in section 5.Results and discussions with Future enhancements are discussed in Section 6 and Section 7 provides the conclusion and future enhancements

II. LITERATURE SURVEY

In spectrographic display, wideband and narrow band analysis methods are used for short and large duration frame analysis. Among these wide band analysis is prominent in frame analysis . The frame analysis is used to achieve high time resolution. If the analysis of high frequency resolutions is required, then larger frames are preferred. The speech signal spectrogram analysis the major roles are played by vowels, consonants, nasals, diphthongs etc, because they carry significant information about the identity of the adjacent character. Normally, it is a common practice to use constant frame size of 20 or 25msec which represents a spectral characteristics of speech sounds[3]. Wide band always plays the major role in computing the dynamics of the speech sounds. Variable frame analysis well supports in capturing the dynamics of the sound. Thus, it is important to compute the dynamics of speech sounds transition well in order to achieve high recognition accuracy. Hence the variable frame numbers

and frame size plays major role in selecting the optimal frames that contains the speech sounds. Some of the existing techniques for fixed and variable frame size is discussed below: The speech signals are quasi-stationary and non-stationary in nature[4] .To analyze these signals generally fixed frame and variable frame length procedures are used. For analyzing vowels and consonants involved speech signals that

are stationary and non stationary for longer durations variable frame analysis procedure is more suitable. Hence these are proposed for speech recognition and speaker identification.

Initially, speech feature vectors (frames) are first extracted at a fixed frame rate the decision for retaining the frames is based on the energy, pitch, frequency, threshold, local maxima etc parameters. But frame analysis has played a major role in selecting the frames in the speech applications. Recent research in VFR analysis moves towards finding optimal representation of a speech signal to improve performance in noisy environments. One of the recent findings is towards the variable frame size which requires the frame length to as small as possible from 25ms to 10ms. This requires frame analysis in steps smaller than the standard 10 ms, while the average frame rate largely remains unchanged. An effective VFR method was proposed that uses a 25 ms frame length with a 2.5 ms frame
shift for calculating Mel-frequency cepstral coefficients (MFCCs) and, conducts frame selection based on an energy weighted cepstral distance[3]. An issue of data compression in distributed speech recognition on the basis of a variable frame rate and length analysis method is proposed in [4]. Energy based frame selection method[6] by delta logarithmic energy using variable frame analysis method is proposed. but
variable frame length is rarely adopted. In [8] the authors focused on speaking rate normalization technique by adjusting frame rate and frame size instead of handling the noise-robustness , is implemented on a state-of-the-art speech recognition architecture and evaluated on the GALE broadcast transcription tasks By using varying number of frames and respective frame lengths of each frame improvement on the speech recognition system can be predicted. The VFL analysis is one of the natural ways of determining frame length.

III. FRAME SELECTION PROCEDURE

In the proposed work the process for optimal frame
selection is adopted by using the thresholding and local
Maximal concepts as discussed below.

a) *Thresholding*: Threshold concept has been applied at every point of framing and windowing. The number of frames considered varies after the deletion of the unwanted frames ie. having lesser energy threshold[7]. Hence the number of frames considered in the processing of MFCC varies before and after the application of energy thresholding concept, hence variable frames.

Frame selection procedure: once the optimal frames are selected they are grouped and interpreted as fuzzy frames. These numbers are selected based on the trial and error procedure assuming more with respect to particular frame the SNR higher frames and lesser the SNR low frames. Since selection of number of frames is not constant hence it is called as ’fuzzy’ interpreted frames.

b) *Local Maxima*

Compute the histogram of the feature sequence value;

- Apply a smoothing filter on the histogram;
- Detect the histogram’s local maxima;
- Let M1 and M2 be the positions of the first and second local maxima respectively; is computed using the following equation

$$T = \frac{W.M1 + M2}{W + 1} \qquad (1)$$

W is a user-defined parameter. Large values of W
obviously lead to threshold values closer to M1 The above process is executed for both feature sequences leading to two thresholds: T1 and T2, based on the energy sequence and the spectral centroid sequence respectively. As long as the two thresholds have been estimated, the two feature sequences are thresholded and the segments are formed by successive frames for which the respective feature values (for both feature sequences) are larger than the computed thresholds.

c) *’fuzzy interpretation*

The first step of both recognition and training processes is to convert the spectrogram of speech signals into a fuzzy frames based on the identification of the information available in each frames[9,18,19]. Generally the fuzzification process is based on four major ideas :

- Unable to extract local features in full precision using spectrograms by human recognizer.
- Rough measure estimates for speech amplitudes by human recognizer .
- Number of speech frames are ignored compared to the lengths of the frames.
- Frames with lower frequency of information is more

Among the four factors the work presented in this papers is towards the third and fourth points used for capturing the dynamics in the speech sounds. The fuzzy interpretation is not assumed at length but at the number of frames(count) . Since relativity is not exact fuzzy interpretation is realized.

d) *Fuzzy interpretation*.

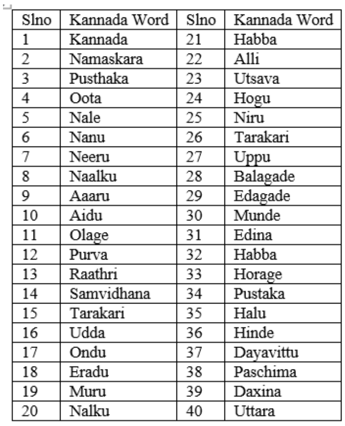
The frames derived from the above procedure are grouped into three levels namely:

- Low frames : In this less number of frames are selected usually towards lower end of the signal .15% of the overall frames are considered as low frames;
- Medium frames 37% of the overall frames are considered towards the middle of the signal.
- High frames: More number of frames are selected i.e 48% of the total frames of the signal are considered towards the beginning of the signal.

IV. DATASET

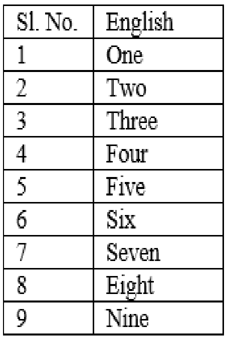
Two different datasets are used for experimentation. The first dataset used is Free Spoken Digit Dataset (FSDD)[10],consisting of recordings of spoken digits sampled at 8kHz. The recordings are trimmed so that they have near minimal silence at the beginnings and ends. It consists of English pronunciations words of numbers from one to nine. Totally 900 signals from four speakers and each digit has total 100 signals as shown in table 2. The second data set used is the Kannada data set of isolated words. The words considered are as shown in table 2. These signals are sampled with16KHz frequency. The dataset consists of 30 speaker’s words, among them 20 male and 10 females. Totally we have 1000 speech samples recorded from both male and female speakers. The noisy data is created by artificially adding 10db babble noise taken from Noizeus database[15,16] to create convoluted noisy speech dataset. The Kannada isolated words are as shown in table 1. Totally kannada words are taken away from users are listed in table 1.

Table 2 Kannada Dataset



Slno	Kannada Word	Slno	Kannada Word
1	Kannada	21	Habba
2	Namaskara	22	Alli
3	Pusthaka	23	Utsava
4	Oota	24	Hogu
5	Nale	25	Niru
6	Nanu	26	Tarakari
7	Neeru	27	Uppu
8	Naalku	28	Balagade
9	Aaaru	29	Edagade
10	Aidu	30	Munde
11	Olage	31	Edina
12	Purva	32	Habba
13	Raathri	33	Horage
14	Samvidhana	34	Pustaka
15	Tarakari	35	Halu
16	Udda	36	Hinde
17	Ondu	37	Dayavittu
18	Eradu	38	Paschima
19	Muru	39	Daxina
20	Nalku	40	Uttara

Table:3 English Dataset



Sl. No.	English
1	One
2	Two
3	Three
4	Four
5	Five
6	Six
7	Seven
8	Eight
9	Nine

V. METHODOLOGY

The speech signals are processed for MFCC features and to classify those features for the speech recognition process. Each frame is processed by applying local maxima and threshold concepts as discussed in section 3.1 and 3.2. The detail procedure for feature extraction and classification are discussed as follows

A. Feature extraction: MFCC – Mel Frequency Cepstrum Coefficients (MFCC) [17,25,26] . Feature

Feature extraction methods are responsible for transforming the speech signal into stream of feature vector coefficients. The obtained fuzzy interpreted optimal frames and their MFCC coefficients are used to discriminate the variations of the speech signal. The framing, windowing, transformation and computation of Mel-cepst are as follows:

a) *Framing: The* given input speech signal is divided into frames of each 20ms. The frames are overlap with1/2 of the frame size. Each frame is processed as discussed in section 3.1 and 3.2 procedures.To obtain optimal frames. 20% of frames are reduced before and after the application of framing process.

b) *Windowing*:
In second step windowing is applied to remove the discontinuities at frame edges. Here each sample of frame is multiplied to hamming widow.

c) Spectral Estimation: Discrete Fourier Transform is applied for each frame to compute spectral coefficients.

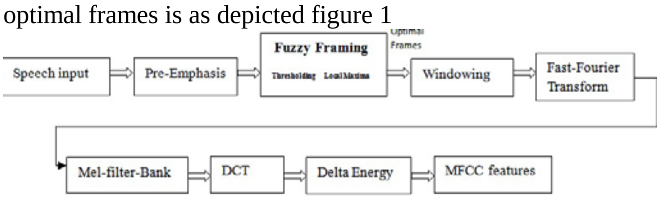
d) Mel filtering: A group of triangle band pass filters are used to simulate the characteristics of the human's auditory system. Mel frequency components are computed using Equation (2).

$$Mel(f) \models \left(1125 * ln(1+\frac{f}{700})\right) \qquad (2)$$

In the final step, the log Mel spectrum is converted back to time domain using DCT. The result obtained is identified as Mel Frequency Cestrum Coefficients (MFCC).

e) *Discrete Cosine Transformation*: It represents the data points , as a sum of cosine functions at different frequencies.

The overall feature extraction procedure retaining the optimal frames is as depicted figure 1



The optimal frames obtained are fuzzy interpreted as High,Mid and Low frames. The fuzzy interpreted optimal frame combinations are as follows: i) High-Mid-Low 53-41-12, ii) High-Low-Mid 53-12-41 iii) Mid-Low-High 41-12-53 iv) Mid-High-Low 41-53-12 v) Low-Mid-High 12-41-53 vi) Low-High-Mid 12-53-41

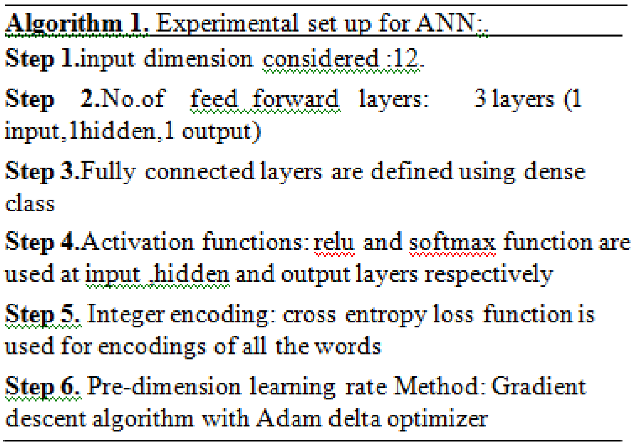
B. Classification.

The classification stage employs an application of various classifiers' like ANN,SVM, kernel SVM and LSTM . LSTM is an variant of RNN. In LSTM each cell able to store information. The LSTM's are used to store temporal dependencies among the data that can be applied for various fields. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that

separates cases of different class labels. An artificial neural network is an interconnected group of nodes, inspired by a simplification of neurons in a brain. The algorithmic steps of applying the modules are discussed in the next section

a) ANN[28,29,30,31,34]:. ANN learning is robust

ANN is robust to errors in the training data and has been successfully applied for learning real-valued, discrete-valued, and vector-valued functions containing problems such as interpreting visual scenes, speech recognition, and learning robot control strategies. The study of artificial neural networks (ANNs) has been inspired in part by the observation that biological learning systems are built of very complex webs of interconnected neurons in brains. Generally, ANNs are built out of a densely interconnected set of simple units, where each unit takes a number of real-valued inputs and produces a single real-valued output.



Algorithm 1. Experimental set up for ANN:.

Step 1.input dimension considered :12.

Step 2. No.of feed forward layers: 3 layers (1 input,1hidden,1 output)

Step 3. Fully connected layers are defined using dense class

Step 4. Activation functions: relu and softmax function are used at input ,hidden and output layers respectively

Step 5. Integer encoding: cross entropy loss function is used for encodings of all the words

Step 6. Pre-dimension learning rate Method: Gradient descent algorithm with Adam delta optimizer

b). *SVM with and Without kernel*:.

A support vector machine (SVM) [11,20]is a machine learning algorithm that analyzes data for classification and regression analysis. In SVM multi-classification is implemented with the "one-against-one" approach (Knerr et al., 1990) . If noofclass is the number of classes, then

noofclass*(noofclass -1)/2

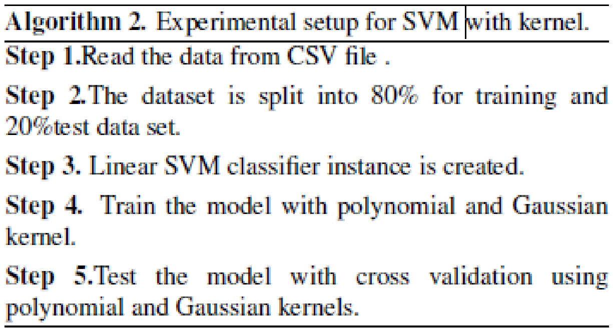
classifiers are constructed and each one trains data from two classes. To provide a consistent interface with other classifiers, the decision function shape option allows to monotically transform the results of the "one-against-one" classifiers to a decision function of shape (n samples, n classes)

Kernels[12]:

The function of the kernel takes nonlinear data as input and transform it into the linear data. Different kernel functions can be applied on SVM . The possible kernels are linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid kernel functions Polynomial kernel The polynomial kernel is commonly used with SVM .It represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models. The polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these samples. This helps in identifying the variations in the speech signal pronunciation that exists during the speech activity. The polynomial and Guassian kernel equations are used in the system design as shown in eq 4 and 5.

$$k(x,y) = \left(\alpha x^{T} y + c\right)^{d} \qquad (4)$$

$$\exp(-\gamma \|x - x'\|^{2}) \qquad (5)$$



Algorithm 2. Experimental setup for SVM with kernel.

Step 1. Read the data from CSV file .

Step 2. The dataset is split into 80% for training and 20%test data set.

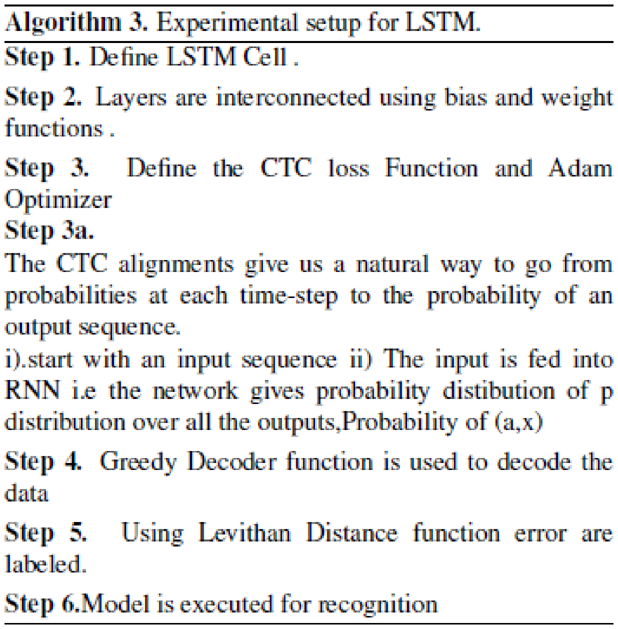
Step 3. Linear SVM classifier instance is created.

Step 4. Train the model with polynomial and Gaussian kernel.

Step 5. Test the model with cross validation using polynomial and Gaussian kernels.

c) *Long short-term memory (LSTM):*

LSTM[21,22,23,24] networks are well-suited in classifying, processing and making predictions based on time series data,there can be lags of unknown duration between important events in a time series. Since speech is dependent on time factor and its variations LSTMs are well suited for this application. These were developed to deal with the exploding and vanishing gradient problems. The figure 2 depicts the training and testing process on all the three classifiers discussed above



Algorithm 3. Experimental setup for LSTM.

Step 1. Define LSTM Cell .

Step 2. Layers are interconnected using bias and weight functions .

Step 3. Define the CTC loss Function and Adam Optimizer

Step 3a.

The CTC alignments give us a natural way to go from probabilities at each time-step to the probability of an output sequence.

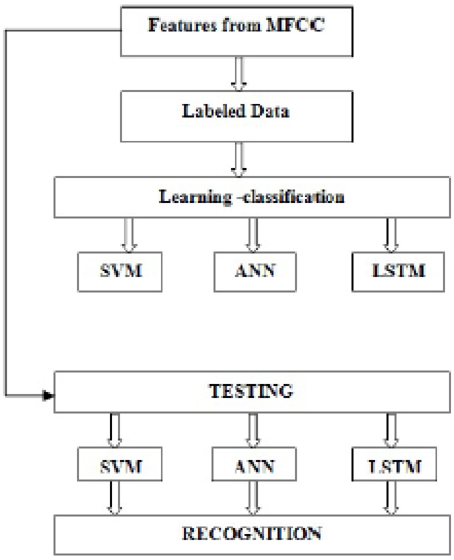
i).start with an input sequence ii) The input is fed into RNN i.e the network gives probability distibution of p distribution over all the outputs,Probability of (a,x)

Step 4. Greedy Decoder function is used to decode the data

Step 5. Using Levithan Distance function error are labeled.

Step 6. Model is executed for recognition

Fig. 2. General Training and Testing Process

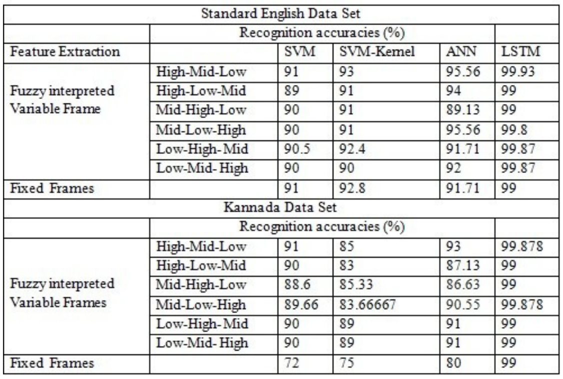


VI. RESULTS AND DISCUSSIONS

The results for fuzzy interpreted variable frames using the above classifiers is presented in the figure 3 and 4. The figure 3 presents the results for optimal frames vs fixed frames used for classification with various classifiers. The classifiers are tested for the various combinations of optimal fuzzy interpreted frames by grouping them in to the terms of high ,mid and low frames for various combinations. The recognition accuracies for all the individual classifiers are tabulated for all the combinations. The obtained recognition accuracy results are compared with fixed frames. Figure 4 presents results available in literature using fuzzy frames with LSTM classifier. Only the Results of LSTM are compared because of the better model availability and for the efficiency of the classifier. From the table 4 it is observed that 2 % WER is presented in the literature. The proposed approach shows that WER still reduction with an improved recognition accuracy up to 99%. But in the literature application of LSTM to Kannada data set in not identified . Hence the same is shown in table. Among various combination ,the fuzzy combination of High-Mid-Low frames(the beginning, medium and the end) has achieved the best accuracies in all the models. But a variation on these combinations sometimes yield us less accuracies also. Since at the beginning the possibility of non-voice signals are predicted more, it requires more frames to be considered. Recognition results for both English and Kannada dataset has been presented in Figure3.

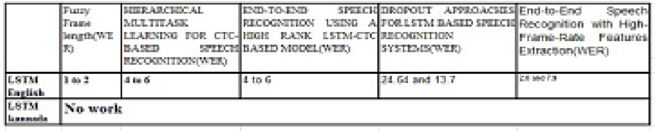
Fig.3.Fixed Frames vs Optimal Frames Average Recognition

.



Standard English Data Set					
	Recognition accuracies (%)				
Feature Extraction		SVM	SVM-Kernel	ANN	LSTM
Fuzzy interpreted Variable Frame	High-Mid-Low	91	93	95.56	99.93
	High-Low-Mid	89	91	94	99
	Mid-High-Low	90	91	89.13	99
	Mid-Low-High	90	91	95.56	99.8
	Low-High-Mid	90.5	92.4	91.71	99.87
	Low-Mid- High	90	90	92	99.87
Fixed Frames		91	92.8	91.71	99
Kannada Data Set					
	Recognition accuracies (%)				
Fuzzy interpreted Variable Frames	High-Mid-Low	91	85	93	99.878
	High-Low-Mid	90	83	87.13	99
	Mid-High-Low	88.6	85.33	86.63	99
	Mid-Low-High	89.66	83.66667	90.55	99.878
	Low-High-Mid	90	89	91	99
	Low-Mid- High	90	89	91	99
Fixed Frames		72	75	80	99

Fig. 4. LSTM Literature Work



	Fuzzy Frame length(WER)	HIERARCHICAL MULTITASK LEARNING FOR CTC-BASED SPEECH RECOGNITION(WER)	END-TO-END SPEECH RECOGNITION USING A HIGH RANK LSTM-CTC BASED MODEL(WER)	DROPOUT APPROACHES FOR LSTM BASED SPEECH RECOGNITION SYSTEMS(WER)	End-to-End Speech Recognition with High-Frame-Rate Features Extraction(WER)
LSTM English	1 to 2	4 to 6	4 to 6	24.64 and 13.7	[illegible]
LSTM kannada	No work				

Discussions or observations on Feature Extraction.

1) Thresholding and local maxima parameters plays a major role in selecting optimal frames when frames are interpreted as fuzzy frames
2) Using Fixed number of frames takes more time and memory yielding the same result.
3) Computational complexity is bit a more for fixed frames compared to variable frames yielding the same results
4) In case of fixed frames all the frames should be processed compulsorily to obtain the same recognition accuracy
5) In the proposed approach since optimal frames are selected the processing of 20%frames are reduced due to the application of thresholding and local maxima concepts.
6) Error rate is still reduced by 1% by adopting optimal frame concepts
7) The vagueness of words exists during pronunciation that are interpreted in terms of frames of High, Mid and Low.
8) some relative lengths are considered frame size also varies
9) Cross validation concept for testing the data suits well when vagueness exists in data.

Figure 5 and 6 shows the graph for fuzzy interpreted classification accuracy for Kannada and English data set. In both data set LSTM performs better compare to figure 3 in results and discussions section.

Observations on classification.

1. Kernel SVM equally works good with LSTM when polynomial RBF kernels are used. .
2. Adam optimzer and CTC loss functions plays the major role in decreasing the WER
3. Generally selection of the normalization methods for SVM plays a major role in varying the results of the models
4. Number of Epochs be decided aptly which plays its

importance in increasing or decreasing the accuracy and error.

5. Among all the classifiers LSTM has highest performance for both the datasets. It is observed that for Kannada data set the recognition accuracy difference is more when compared to fixed and variable frames. SVM with polynomial kernel and the LSTM models equivalently performs good.

Fig. 5. Graph for average recognition accuracy for English Data.

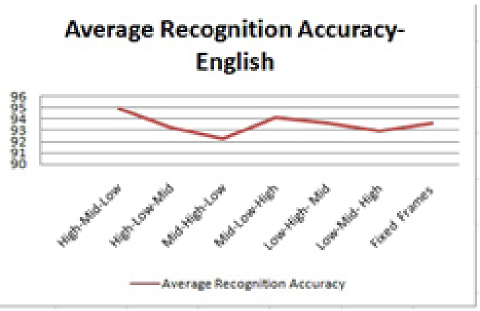
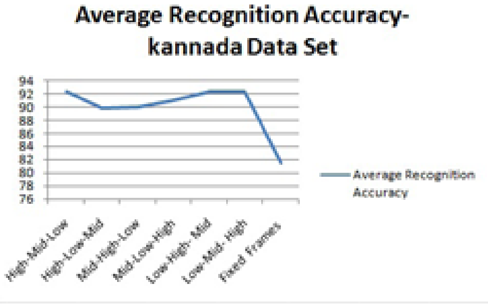


Fig. 6. Graph for average recognition accuracy for Kannada Data.



VII. CONCLUSIONS AND FUTURE ENHANCEMENTS

It is observed that in fixed frame analysis all the frames compulsorily need to be processed. From the proposed approach its observed that 20% of the frame processing is reduced because fuzzy interpreted ,local maxima and Thresholding concepts. The fuzzy interpretation helps in identifying the details that make the recognition slow and sensitive to small pertuburation or noise. In both the data sets LSTM model with proposed approach performs better by increasing 1.25% of recognition accuracy compare to fixed frames. 5% For Kannada data set and increase of 11.71% of recognition accuracy compare to fixed frames is observed has been identified. Finally LSTM with variable number of frames can be used to increase the recognition rate. The work can be further enhanced by considering the following:

1) The work can be further enhanced by considering :By reducing the frame length from 20 msec to 15 or 10 msec and then apply the combination of HLM.
2) The performance of the models can be verified for various types of noise and levels
3) The models optimal frames and their fuzzy interpretation can be validated for sentence level and connected word recognition

ACKNOWLEDGMENT

The authors remain grateful for all the persons who have helped us in understanding and formulating this paper. We are very much thankful to Dr.S.K.Katti, for making us to realize the basic mathematical concepts behind the study.

REFERENCES

[1] L. Rabiner and B.H. Juang, “Fundamentals of Speech Recognition” ,Prentice-Hall, ISBN 0-13-015157-2.
[2] Zheng-Hua Tan ,Ivan Kraljevski , “Joint variable frame rate and length analysis for speech recognition under adverse conditions”, Computers and Electrical Engineering 2139–2149 2014:
[3] Samudravijaya K, “Variable Frame size analysis for speech recognition”, http ://www .iitg.ac.in /samudravijaya /publ/04icon VframeSize.pdf..
[4] Ivan Kraljevski1 , Zheng-Hua Tan, “VARIABLE FRAME RATE AND LENGTH ANALYSIS FOR DATA COMPRESSION IN DISTRIBUTED SPEECH RECOGNITION”, Proceedings of IC-NIDC,2014..
[5] E C-S. Jung, KJ. Han, H. Seo, S.S. Narayanan, and H.G. Kang , “A variable frame length and rate algorithm based on the spectral Kurtosis measure for speaker verification” , INTERSPEECH, Makuhari, Japan, International Secretariat, Institute of Pacific Relations, September 2010
[6] Julien Epps Eric H. C. Choi, “An energy search approach to variable frame rate front-end processing for robust ASR”, INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005.
[7] Theodoros Giannakopoulos, “ A method for silence removal and segmentation of speech signals, implemented in Matlab” , University of Athens, Athens, 2009,Volume 2.
[8] M. J. F. Gales et al, “ Support Vector Machines for Noise Robust ASR”, http:// citeseerx.ist .psu. edu/viewdoc/ download?doi=10.1.1.297.7857.
[9] Ramin Halavat et al, “ Recognition of human speech phonemes using a novel fuzzy approach” , Artificial Intelligence Lab 308, Computer Engineering Department,Sharif University of Technology, Tehran, Iran, 2007
[10] https://github.com/Jakobovski/free-spoken-digit-dataset, : DOI 10 .5281 / zenodo.1342401
[11] Osman Eray ; Sezai Tokat ; Serdar Iplikci , “An application of speech recognition with support vector machines” , 6th International Symposium on Digital Forensic and Security (ISDFS), 10.1109, 2018
[12] Aravind Ganapathiraju et al , “Support Vector Machines for Speech Recognition” ,http://citeseerx. ist.psu.edu /viewdoc /download ?doi=10.1.1.131.6537
[13] DAVIS, K. H,et al, “Automatic Recognition of Spoken Digits” , The Journal of the Acoustical Society of America, 1952
[14] KATAGIRI, S , “Pattern Recognition in Speech and Language Processing”, CRC Press 2003.
[15] Hu, Y. and Loizou, P Subjective evaluation and comparison of speech enhancement algorithms, Speech Communication, 49, 588-601,2007
[16] Hu, Y. and Loizou, P Evaluation of objective quality measures for speech enhancement IEEE Transactions on Speech and Audio Processing,16(1), 229-238;2008.
[17] M A Anusuya and S K Katti , “Speech Recognition by Machine: A Review”, International Journal of Computer Science and Information Security Vol. 6, No. 3:2009
[18] Patrick M. Mills Fuzzy Speech Recognition.
[19] Ramin Halavati , “Evolution of Speech Recognizer Agents by Artificial Life” , World Academy of Science, Engineering and Technology 11 2005
[20] Ruben Solera-Ure˜na et.al , “ SVMs for Automatic Speech Recognition,A Survey” , LNCS 4391, Springer-Verlag,Workshop on Nonlinear Speech Processing (WNSP 2005)
[21] Yangyang Shi Mei-Yuh Hwang Xin Lei “, End-To- End Speech Recognition Using a High Rank LSTM-CTC Based Model http://ssli.ee.washington.edu/ mhwang/mobvoi:2019

[22] Jayadev Billa , "Dropout approaches for LSTM based speech recognition systems, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP):2018

[23] Kalpesh Krishna et.al Hierarchical Multitask Learning for CTC-BASED Speech recognition , arXiv:1807.06234

[24] Yan-Hui Tu , "A Hybrid Approach to combining conventional and deep learning techniques for single channel speech enhancement and Recognition" , IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2018

[25] Vani H Y AnusuyaMA ,Isolated Speech recognition using K-means and FCM Technique International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT),2015.

[26] Vani H Y Anusuya M A,Noisy speech recognition using KFCM International Conference on Cognitive Computing Information Processing:2017 .

[27] Jeen-Shing Wang,et al , "A Cluster Validity Measure With Outlier Detection for Support Vector Clustering IEEE Transactions on Systems, MAN, And Cybernetics-Part B: Cybernetics VOL. 38, NO.2008

[28] Salmela, P,, "Neural Networks in Speech Recognition Tampere University of Technology, Publications 295, 2000

[29] Blankenstein, B , "Artificial Neural Networks (ANN)", Department of CS,Washington University in St Louis, 2001

[30] Hosom, J. P et al , "Speech Recognition Using Neural Networks" at the Center for Spoken Language Understanding ,Turorail 1999

[31] Veera Ala-Keturi , "Speech Recognition Based on Artificial Neural Network" , Helsinki University of Technolog

[32] D. Tran, M. Wagner , Fuzzy hidden Markov models for speech and speaker recognition NAFIPS, 1999, pp.426–430.

[33] L.A. Zadeh , From computing with numbers, to computing with words, a new paradigm Int. J. Appl. Math. 12 (3) (2002),307–324

[34] Kurogi, S ,Speech recognition by an artificial neural network using findings on the afferent auditory system. Biological Cybernetics, 64(3), 243–249,1991