

# Experimental Analysis on Performance of Speech Utterance recognition using AI Models

Srikanth G N<sup>1</sup>, M K Venkatesha<sup>2</sup>

<sup>1</sup>Department of Electronics and Instrumentation Engineering, RNS Institute of Technology, VTU, India

<sup>2</sup>Department of Electronics and Communication Engineering, RNS Institute of Technology, VTU, India

**Abstract:-** In Automated speech recognition of the system performance is crucial and important to satisfy multiple requirements of HMI and, more recently, even in IoT-related applications as well. Concurrently, there has been an increase in demand for detecting strong critical features derived from speech utterances. This paper presents a performance of the developed machine learning algorithms with respect to audio digit speech recognition and classification. The prepared dataset contains a free range of words (from 1 to 10) from speakers of different age groups. The Audacity software used for preprocessing the audio files that includes removal of noise included in the signal and trimming the silence on either side of the word utterance. audio signal sampled at  $f_s = 48\text{kHz}$ . We have developed four AI Models to recognise the word utterances. Audio signals are processed separately and derived two unique feature sets that includes statistical features set and singular values by performing SVD related to word utterances. The cepstral values for each utterance are obtained from state-of-the-art MFCC. Variance-covariance matrix is calculated from the generated MFCC matrix. The diagonal values which form the variance are recorded and denoted as feature set-1 for the word utterance and inputted to the machine learning algorithms. Performance matrices of the developed models are recorded. To keep the computational bottleneck associated with the use of feature sets to minimum, dimensionality reduction is carried out by applying singular value decomposition to the extracted MFCC matrix. The derived set of singular values considered as feature set-2 is used to train and test the developed AI models with a ratio of 70:30. We presented and discussed the performance and results produced by MLP, KNN, SVM, Random Forest algorithms. In comparison, MLP and Random Forest were found to show excellent performance on both feature sets with 100% training accuracy and 99% test accuracy.

**Keywords-** Human machine interface (HMI), Mel frequency cepstral coefficient (MFCC), Singular value decomposition (SVD), Multilayer Perceptron, Random Forest.

## 1. Introduction

Speech is a major method of communication to convey the information or the internal feelings of human sometimes more than the text. Speech Signal is concerned with improving speech quality, especially for HMI applications. Automated speech recognition is studied and investigated by researchers for many years with the goal of realizing machines/robots that can recognize speech and follow instructions or commands. Comparison of speech waveforms is often very difficult with respect to amplitude variations and even the phase can change according to transmission and recording systems, it is desirable to remove the phase components from speech wave. Therefore, short time spectral density is usually extracted at short intervals for analysis. It is known that during these short interval's speech is considered to be statistically stationary.

To facilitate Human Machine Interface, the system must be designed to be user friendly and must exhibit higher degree of accuracy. Speech recognition techniques involve speech signal pre-processing, filtering, sampling and applying transformations to suppress or remove the redundant and unwanted part buried in signal of interest while preserving the information content. Though many methods available for extracting speech features, the most

widely used method is to extract MFCC (Mel Frequency Cepstral Coefficients). [4]-[7], which is state of the art because it mimics human speech perception. If we need to track speech dynamics with cepstral coefficients, for example trajectory of MFCC coefficients over time. We can also determine the velocity coefficient (derivative  $\delta$ ) and the acceleration coefficient however computation complexity will increase. Moreover, to extract the main features of speech and reduce the computational complexity, we have to apply dimension reduction techniques like, differential transformation matrices or PCA (Principal Component Analysis) or SVD (singular value decomposition) to obtain significant features.

For audio isolated word recognition, where the wavelet transform is used for feature extraction, it reduces the computational complexity of the neural network by minimizing the utterance of word into a lower-dimensional feature vector [11]. Specifically, to distinguish spoken words or digits, we need to obtain statistical information associated with the signal to serve the purpose. The extracted salient features are labelled acoustic feature vectors fed as input to a ML algorithm and labelled known data considered output. In this work, we implemented four different algorithms i) Random Forest ii) SVM iii) KNN and iv) MLP. However, Deep Neural Network requires a large training dataset to perform classification and prediction. For a multi-class problem like in our work, considering the volume of the database, MLP proved to be the most appropriate classification technique intended for limited dataset records. This work finds its applications in areas such as HMI, IoT, Robotics and speech recognition engines etc.

## 2. Objectives

ASR systems have principally focused on phoneme, word, hence sentence decoding and identification of speaker using various algorithms and techniques like LPC [8] [13] with Hidden Markov Model [14][16][18][20] where they predict the output based on expectation maximization by reducing error. The ASR systems and speaker identification application includes auto correlation analysis and LPC analysis [8],[13],[12], it is revealed that features extracted using MFCC perform better compared to LPC [9] in speech recognition. With regard to acoustic feature extraction, researchers used Mel Frequency Cepstral Coefficients, since cepstral coefficients mimics human perception [2],[8],[4],[5],[10],[13]. In ASR systems, Hidden Markov model initially used in the prediction of speech utterances [14], where in the phoneme is considered the basic unit of speech, the combination of phonemes forms the word utterance. HMM states are generated for a word conferring to the transition probabilities during training process. In the test cycle, for the given input, the probability that a target sequence is generated from each word vocabulary is calculated, and identification is made based on the highest accumulated probability value. Traditionally, ASR systems been developed using HMM with Gaussian Mixture output distribution, but have been found to produce diminishing returns and also accuracy [17], mainly due to fact that the training phase is complicated and computationally demanding, and any misalignment of the states, which can cause a malfunction in the recognition phase [14],[16],[18]. Applying machine learning algorithms found to be a replacement for HMMs to improve the performance of ASR systems, as it holds the promise of substantially improving ASR generation by rigorously checking the efficiency of advanced strategies [1],[3],[4].

The researchers applied different machine learning algorithms/Models and found markable success in classification of spoken speech signals as applied to speech detection and also speaker identification systems with considerable accuracy ranging from 70 to 98% depending on the language used and the algorithms. Methods and algorithms that can be followed for speech processing are well described in [19].

The set Objectives are:

1. To extract significant features, specifically statistical features, from audio speech utterances.
2. To develop AI models for recognition and classification of word utterances.
3. Compare the performance of developed AI models

## 3. Proposed Methodology

In principle, most ASR systems aimed to incorporate a feature selection and extraction block as the main module. In the course of feature extraction, short segments of speech (30-40 ms) are taken from the utterance of a word in sequence. The vocal tract shape almost remains constant hence in this period speech is considered to be statistically stationary. The feature extraction module is constructed to extract the significant features that

ultimately aid the word recognition. The speech features are derived using the MFCC model. Here, the system aims to convert the speech wave into a parameter shape for further investigation, as mentioned, we are considering short segments of speech, so due to the restricted number of samples, we will get reliable spectral data, but with longer duration, speech characteristics will change, therefore the short-time Fourier transform is considered for speech analysis in the spectral domain. The window that can be used due to its acceptable characteristics will be Hamming window because it has good sidelobe attenuation. A window overlapping of 25%, which corresponds to 10ms, provides a smooth transition and prevents spectral distortion. Windowing operation is expressed in (1). If  $w(n)$  represents a window defined for  $0 \leq n \leq N-1$ , where  $N$  denotes total samples in each segment, then the windowing process results in

$$y(n) = x(n) \cdot w(n) \quad 0 \leq n \leq N-1 \quad (1)$$

Hamming window formed from the cosine term is showed in Equation (2).

$$w(n) = 0.54 - 0.46 \cos[2\pi n/(N-1)] \quad 0 \leq n \leq N-1 \quad (2)$$

Here,  $N$  represents total samples in the block. The phases used feature extraction, recognition and classification of digit utterance is portrayed in Figure 1. and Figure 2.

### 3.1 Mel Frequency Cepstral Coefficients:

The steps for obtaining the MFCC coefficients for speech is represented in Figure 1. It is known that human generated speech will be in around 4 kHz range, which includes almost all the human speech sound energy, so a typical sampling frequency of 8 kHz is sufficient. In our experiments, the sampling frequency is considered to be 48 kHz, since it is compatible with most of today's modern machine interface systems. The MFCC is chosen because it mimics the functional characteristics of human ears and sound perception, i.e., MFCC has a linear response up to 1000 Hz and becomes logarithmic after 1 kHz.

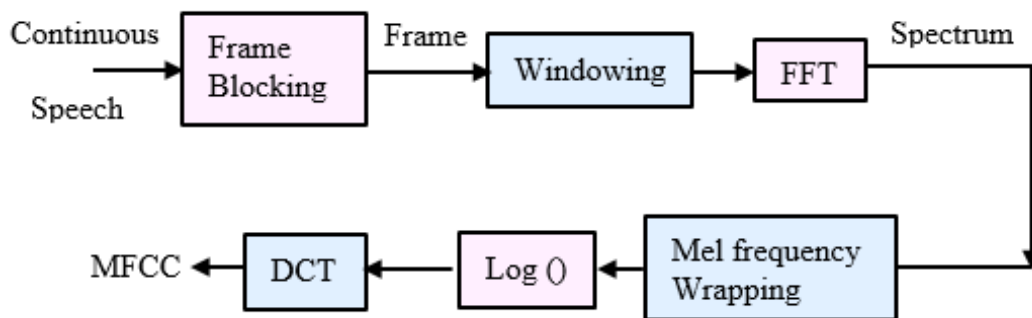


Figure 1. MFCC from continuous speech signal

MFCC Procedure involves following steps

1. Segments of speech from a full speech consecutively and process each segment to obtain reliable features.
2. FFT is applied to windowed signal to convert each segment of  $N$  samples to spectral domain to examine the spectral content.
3. For each individual tone signal with real frequency  $f$ , a subjective pitch or frequency is determined using the "Mel" scale.
4. The equation for frequency mapping from linear scale to Mel scale is expressed in (3).

$$M(f) = 1125 \ln \left( 1 + \frac{f}{700} \right) \quad (3)$$

5. To obtain the original frequency, the inverse equation is given by (4).

$$M^{-1}(m) = 700 \left( \exp \left( \frac{m}{1125} \right) - 1 \right) \quad (4)$$

Mel scale is a linear frequency scale below 1 kHz and a logarithmic scale above 1 kHz. A way to mimic a Mel spectrum is to use a filter bank, which is a sequence of triangular BPFs, individual filter being in a filter bank and center frequencies evenly spaced on the Mel scale. The space between the center frequencies and the bandwidth is determined by a constant interval of Mel frequencies. To calculate the strength of a filter group, we multiply each filter group by power spectrum and then add the coefficient.

- The MFCC set is obtained by applying DCT. DCT is needed to decorrelate the filter bank coefficients. Denoting those Mel coefficients of power spectrum resulting from the last step as

$$\tilde{S}_k, k = 1, 2, 3, \dots, K \quad \text{we can calculate MFCC i.e., } \tilde{c}_n \text{ is expressed in (5)}$$

$$\tilde{c}_n = \sum_{k=1}^K (\log \tilde{S}_k) \cos \left[ n \left( k - \frac{1}{2} \right) \frac{\pi}{K} \right] \quad (5)$$

In each frame, the sound feature vector consists a total 13 selected real scalars, we achieved quite good results using only the selected number of cepstral coefficients and we wanted to keep number of elements as low as possible to lessen computational complexity, velocity and acceleration coefficients, hence lower dimension space.

### 3.2 Feature extraction:

#### 3.2.1 Statistical Variance extraction

The first element is omitted from the generated MFCC coefficients because it contains information related to the energy content rather than the configuration of the vocal tract and thus the acoustic vector sequence, which forms a matrix of MFCC coefficients/features with 13 columns and the number of occupied rows depending on the length of the statement.

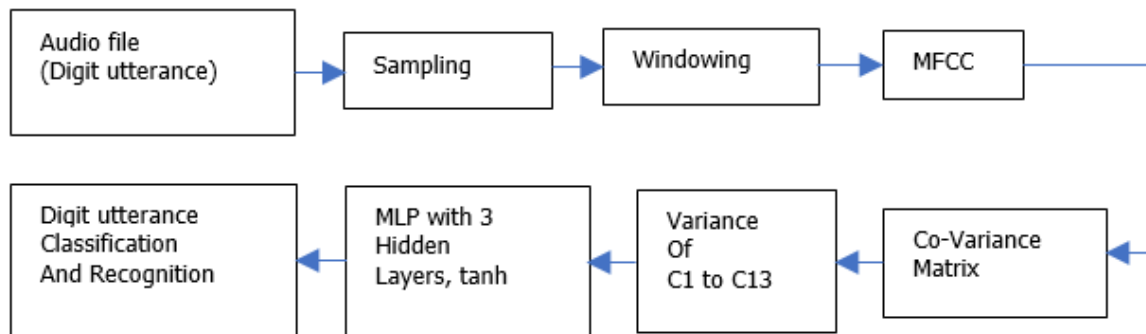


Figure 2. Steps followed in Digit Utterance Recognition

Steps followed Digit utterance classification and recognition are

- To obtain its statistical variations on selected variables, a covariance matrix is created from the obtained MFCC matrix.
- A general representation of covariance considering two random variables is shown in equation (6).

$$\text{COV}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - E(X))(y_i - E(Y)) \quad (6)$$

- The statistical variance is obtained by extracting the diagonal elements arranged in descending order to form a feature vector from the covariance matrix.
- The post-processed feature vector is referred to here as MFCC and is tabulated in a .csv file indicating the acoustic vector input variables against the labelled known output. Figure 3 depicts the sample of variance features C1 to C13 for digit utterance 3 and 5.
- Figure 2 depicts the steps followed for digit word utterance recognition and classification.
- 6a. Apply MLP to features from step4 with 3 hidden layers with 'tanh' activation function and hyper parameter tuning.

- 6b. Get AI model performance metric.
7. Repeat step 6 for KNN, SVM, Random Forest algorithm
8. Output: Evaluation of classification results and comparing the AI model performance.

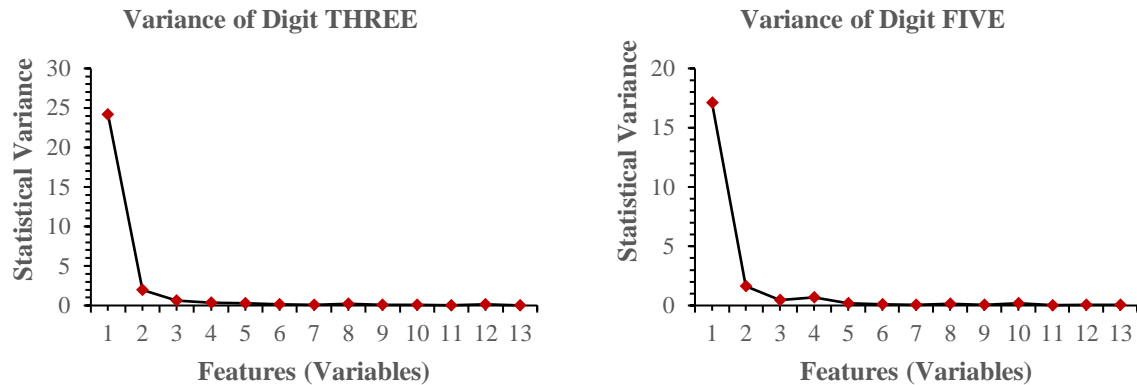


Figure 3. Statistical Variance features for digit utterance 3 and 5

### 3.2.2 SVD feature extraction

We have prepared another feature set for the dataset considered by applying singular value decomposition on over extended matrix and this basically minimise the dimensionality of the problem. SVD will decompose the input matrix into set of orthogonal matrices and a diagonal matrix, and the equation is shown in (7)

$$A=U S V^T \quad (7)$$

Here 'A' is the MFCC matrix. 'U' is an orthonormal matrix (matrix of  $m \times n$  orthonormal eigenvectors of  $AA^T$ ). 'V' is an orthonormal matrix (the transpose of an  $n \times n$  matrix containing the orthogonal eigenvectors of  $A^T A$ ). 'S' is a diagonal matrix containing the singular values arranged in descending order to form acoustic vector. The singular values are the square root of the eigenvalues. SVD decomposes the data matrix (not necessarily a square matrix) into a low-level matrix. A further approximation is obtained by keeping an optimal number of singular values that are arranged in descending order ( $\sigma_1 > \sigma_2 > \sigma_3 \dots > \sigma_p$ ) and removing all redundant values smaller than  $\sigma_p$  which contains very less information. A set of features obtained from SVD analysis was found to enhance the accuracy of word recognition. We treated the singular values as a data matrix to reduce the computational complexity and populated them into a .csv file containing the labelled inputs and output for the entire selected data set. Figure 4 displays the sample of the SVD singular values obtained for digit utterance 3 and 7.

Steps for audio digit recognition using MLP and KNN on SVM, Random Forest algorithm on SVD features

1. Input: Digit Utterance Audio .mp4 file
2. Audio signal sampling with  $fs=48\text{kHz}$ .
3. Window using Hamming window  $w(n)$
4. Generate the MFCC coefficient matrix
5. Apply SVD on the overextended MFCC matrix
6. Extract the singular values of chosen variables and arrange in descending order and populate the obtained acoustic feature in .csv file
7. 7a. Apply MLP to features from step 6, with three hidden layers and an activation function tanh, hyperparameter tuning.
- 7b. Get AI model performance metrics

8. Repeat step 7 for KNN, SVM, Random Forest algorithm
9. Output: Evaluation of classification results and comparing the AI model performance.

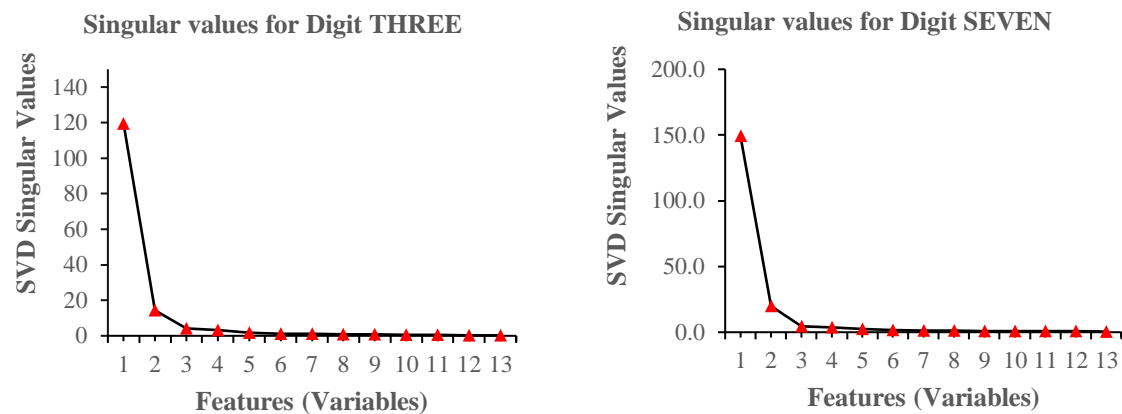


Figure 4. SVD features for digit utterance 3 and 7

### 3.3 Proposed Algorithms

#### 3.3.1 Random Forest Classifier

Random forests are a type of ML algorithm used in classification and regression tasks. It works on the principle of creating decision trees considering random subset of data. The concept followed is to aggregate the prediction result from multiple decision trees and produce a final result that is based on an averaging mechanism. Therefore, it is less prone to overfitting. Random Forest tries to maximize the information gain at each node that is decreasing entropy, which is seen as increasing the purity of the node. However, it is quite difficult to identify which variable contributes to higher information gain, and considerably gives good results because we decorate the trees at the expense of slow training time.

In our work, "entropy" is used as the criterion instead of "gini", and by tuning the parameters through experimentation, we were able to obtain a substantial improvement in accuracy. Using 1st feature set formed by taking statistical variance on MFCC covariance against C1 to C10 (by dropping coefficients C11 – C13), with training and testing ratio of 75:25 and considering equal no of records for testing we have got training accuracy of 100% and testing accuracy of 95.4%, with mse of 0.045.

The same model when used considering 2<sup>nd</sup> feature set formed from SVD to extract singular values related to word utterances as shown in figure 4, we have obtained training accuracy of 100% and testing accuracy of 99% with a mean square error of 0.034.

#### 3.3.2 SVM Classifier

Support Vector Machines (SVMs) are learning models that can be used for both classification and regression tasks. Uses a subset of training points in the decision function (called support vectors), it is memory efficient. SVMs work by mapping data points into a higher-dimensional feature space, which lets to capture non-linear relationships between features and perform complex classifications and regressions. SVM is based on statistical approaches. We have used SVM model with a C value of 1.5, linear kernel and using 1st feature set C1 to C10 (by dropping coefficients C11 – C13), with train and test ratio of 80:20 and considering equal no of records for testing we have got training accuracy of 98.9% and testing accuracy of 92.8%, with a mean square error of 0.072. The same model when used considering 2<sup>nd</sup> feature (SVD) we have obtained training accuracy of 98.6% and testing accuracy of 97% with a mean square error of 0.032.

#### 3.3.3 KNN Classifier

K-Nearest Neighbor is a supervised machine learning class, the data to be predicted is assigned to a cluster of centroids, based on the similarity measure i.e., the distance function. KNN provides highly adaptive behaviour



and optimal in the large sample limit. The disadvantage is that it is computationally intensive and requires a lot of storage since it has to memorise the trained data.

KNN is a traditional classification method and requires no training effort, it strongly depends on the quality of measurements between samples, noise can easily affect the performance of this classifier. The distance function used here is Euclidean distance, shown in Equation (7), which is useful in a low-dimensional dataset.

$$d_{\text{euclidean}}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (7)$$

where  $x$  and  $y$  are  $n$ -dimensional vectors and denoted by  $x = (x_1, x_2, x_3 \dots x_n)$ ,  $y = (y_1, y_2, y_3 \dots y_n)$  represent  $n$  attributes of two records. We also used a KNN classifier for digit utterance recognition on the features set 1 and 2 related to word utterances.

Using KNN on feature set1, we achieved a training accuracy of 86% and a testing accuracy of 78% with a chosen training to testing ratio of 80:20 and considering the equal number of records. Model is found to exhibit overfit for the selected set of features, although the model performance is checked for different values of "k" and features were also selected that has provided good accuracy in other models. Weight and noise can be factors that affect model performance. Using the same model "w.r.t". feature set 2 (SVD), we observed a significant increase in F1 scores noticeable in Table 3 and increase in accuracy of 15.4%, depicted in Table 5. We obtained a training accuracy of 97.8% and a testing accuracy of 93.4% with a root mean square error of 0.097, and the observed deviation in accuracy between the two phases is 4.4 %, hence optimally a best fit.

### 3.3.4 Multilayer Perceptron Classifier (MLP)

Multilayer perceptron is a network consisting of an input layer, one or more hidden layers, and an output layer. It is essentially a feedforward ANN with a supervised learning strategy and using in backpropagation methods, the network weights are tuned to minimize loss function and achieve convergence [5][16], thereby driving the prediction error to zero. Multiple layers of MLP with non-linear activation function like tanh and ReLU differentiate MLP from linear perceptron.

Through experimentation, we found that the rectified linear activation function gives promising results. ReLU is a piecewise linear function, if the input is positive then the function outputs the input value itself, otherwise the output will be zero, ReLU is used in hidden layers of MLP. In addition, we also found that tanh as an activation function showed improved accuracy compared to ReLU, since the derivatives are not monotonic, tanh solves the dead neurons problem, which is requirement in any backpropagation to diminish the error and hence improve accuracy, the results shown show MLP using the tanh activation function. However, the use of DNN requires large database records for training. In [10] we have experimented using Social Exponential Optimization Algorithm Based Deep Residual Network for Visual Speech Recognition.

For the above mentioned model implementations, MATLAB is used to develop programs for speech processing program, to generate MFCC coefficients and extract statistical deviations of significant sign with respect to pronunciation of digits. The obtained feature set 1(Variance) and features set 2(Singular values) are separately populated in .csv files. Program is written in python to input the acoustic vectors to MLP. Considering a training to testing ratio of 70:30. For the 1st feature set, we have obtained a training accuracy of 100% and a testing accuracy of 99% with mean square error of 0.0096.

The same model when used considering 2nd feature set (SVD) we have obtained training accuracy of 100% and testing accuracy of 99% with a mean square error of 0.0076. It is clear that MLP perform extremely well, confusion matrix and classification report is portrayed in table 3 and table 4.

## 4. Experiments

### 4.1 Dataset

In our models proposed we have used a dataset of spoken speech digits. The data collection consists of 345 audio files in .mp4 format consisting of 36 subjects including men and women of different age groups (excluding children) speaking in ten sentences. The sentence contains utterances with numbers from one to ten, 358 files from which the spoken utterances are extracted using "Audacity", an audio processing software tool.

## 4.2 Experimental results and discussions

In our experiments we have developed four AI Models, Random Forest, SVM, KNN and MLP for recognising the digit utterances. All the mentioned models are trained and tested for its performance on the constructed feature sets 1 and 2 as mentioned in section 3.2.1 and 3.2.2.

### 4.2.1 Considering Feature set 1 (Variance)

We observed that considerably good performance in all the developed models except KNN which has recorded an accuracy of 78% slightly overfit due noise and hence outliers. It is also known that KNN memorises the training phase records to identify the test case. Table 1 exhibits the F1 score of the models w.r.t digit utterance. Table 2 depicts the accuracy attained in training and testing phase of the developed models. Figure 6 shows the loss observed against epochs and hence convergence obtained. Figure 5 exhibits the Performance Assessment of MLP, Random Forest, KNN, SVM Algorithms on Variance Features. From the obtained classification and accuracy, it is known that MLP and Random Forest superior performance compared to SVM and KNN.

Table 1. Performance metric F1 score obtained from models for Digit Utterance Recognition on Variance Feature Set.

DIGIT DATA	F1-Score of the Models on Variance Feature set			
	Random Forest	SVM	KNN	MLP
1	82	80	80	96
2	84	77	77	95
3	100	100	93	100
4	100	100	100	100
5	100	100	71	100
6	100	100	75	100
7	94	100	73	100
8	94	83	62	100
9	100	88	67	100
10	100	100	83	100

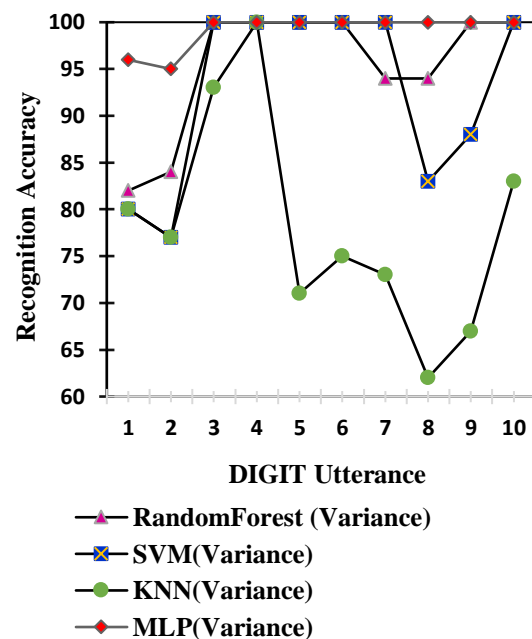


Figure 5. Accuracy Performance Assessment of MLP, Random Forest, KNN, SVM Algorithms on Variance Features

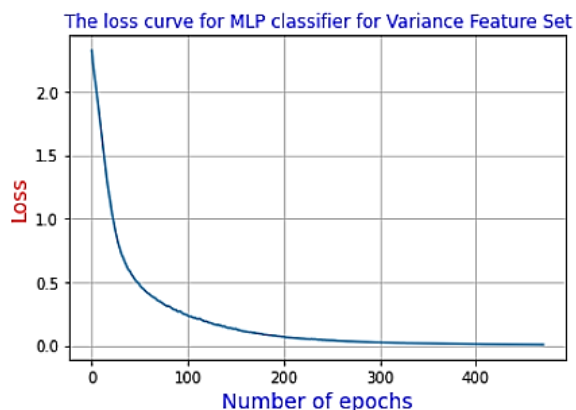


Figure 6. Loss against number of epochs in MLP performance on Variance feature set

Table 2. Performance of Classifiers Applied to Digit Utterance Recognition on Variance Feature Set.

Digit Data	Classifier	Recognition Accuracy in %		MSE
		Training Accuracy	Testing Accuracy	
1-10	Random Forest	100	95.4	0.045
1-10	SVM	98.9	92.8	0.072
1-10	KNN	86	78	0.463
1-10	MLP	100	99	0.009



#### 4.2.2 Considering Feature set 2 (SVD)

The confusion matrix shown in Table 3 exhibits the uttered digit recognition against the test samples using MLP classifier. The principal diagonal elements depict the degree of predicted output against the actual class. Although considerably good level of classification is observed in all the testing cases, however mis classification is seen in test case for the digit 6 with testing cases 13. Table 4 displays the detailed classification report wherein the obtained f1score can effectively be considered against the chosen dataset. Figure 8 shows the loss curve against the epochs and convergence. Table 5 shows F1 score obtained from models for Digit Utterance Recognition on SVD Feature set. Figure 7 portrays Performance of MLP, Random Forest, KNN, SVM Algorithms on SVD Features. Table 6 depicts the accuracy attained in training and testing phase of the developed models. On comparison between the accuracy attained by using Random Forest, KNN, SVM and MLP algorithms. The developed AI models exhibited good performance using feature set 2 compared to feature set 1, since singular value decomposition, will decompose the given over extended MFCC feature matrix into orthogonal set that forms independent unique vectors along with greatest set of singular values.

Table 3. Confusion Matrix for Digit Classification Using MLP on SVD features

	PREDICTED BY MODEL									
	1	2	3	4	5	6	7	8	9	10
1	14	0	0	0	0	0	0	0	0	0
2	0	14	0	0	0	0	0	0	0	0
3	0	0	14	0	0	0	0	0	0	0
4	0	0	0	14	0	0	0	0	0	0
5	0	0	0	0	14	0	0	0	0	0
6	0	0	0	0	0	12	1	0	0	0
7	0	0	0	0	0	0	14	0	0	0
8	0	0	0	0	0	0	0	13	0	0
9	0	0	0	0	0	0	0	0	14	0
10	0	0	0	0	0	0	0	0	0	14

Table 4. Classification-Report for Digit Utterance Recognition Using MLP on SVD features.

Metric Class	Precision	Recall	f1-score	Support
1	1.00	1.00	1.00	14
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	14
4	1.00	1.00	1.00	14
5	1.00	1.00	1.00	14
6	1.00	0.92	0.96	13
7	0.93	1.00	0.97	14
8	1.00	1.00	1.00	13
9	1.00	1.00	1.00	14
10	1.00	1.00	1.00	14
accuracy			<b>0.99</b>	138
macro avg	0.99	0.99	0.99	138
weighted avg	0.99	0.99	0.99	138

Table 5. Performance metric F1 score obtained from models for Digit Utterance Recognition on SVD Feature Set.

DIGIT DATA	F1-Score of the Models on SVD Feature set			
	Random Forest	SVM	KNN	MLP
1	100	100	100	100
2	100	100	91	100
3	96	100	88	100
4	100	100	100	100
5	96	100	90	100
6	100	100	88	96
7	100	88	95	97
8	100	90	100	100
9	100	95	89	100
10	100	94	94	100

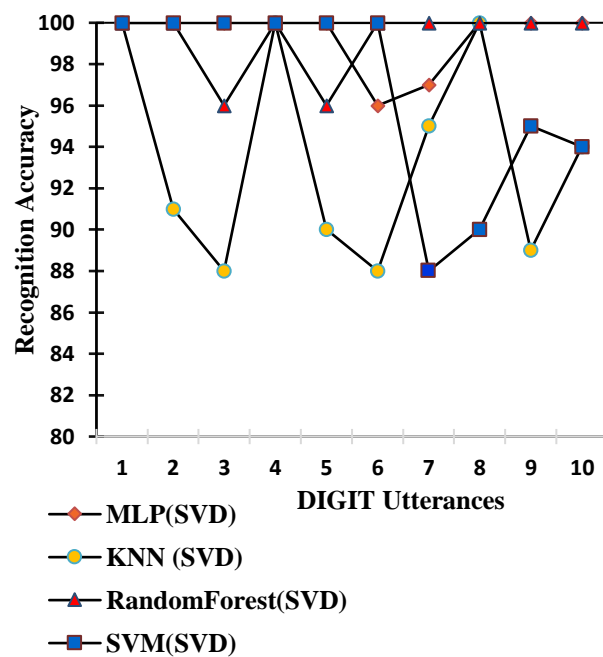


Figure 7. Accuracy Performance Assessment of MLP, Random Forest, KNN, SVM Algorithms on SVD Features

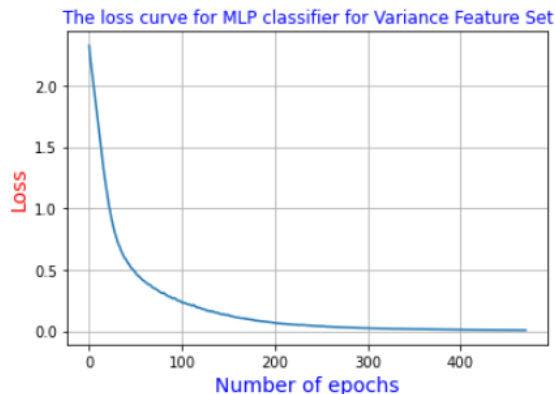


Figure 8. Loss against number of epochs in MLP performance on SVD feature set

Table 6. Performance of the Classifiers Applied to Digit Utterance Recognition on SVD Feature Set.

Digit Data	Classifier	Recognition Accuracy in %		MSE
		Training Accuracy	Testing Accuracy	
1-10	Random Forest	100	99	0.034
1-10	SVM	98.6	97	0.032
1-10	KNN	97.8	93.4	0.097
1-10	MLP	100	99	0.007

## 5. Discussion

In this work we have shown that the digit utterances are recognised and classified with an accuracy of 99%. The experimentation involves considering statistical variance as extracted significant feature from the audio digit utterance. we have used four classification algorithms namely Random Forest, SVM, KNN and MLP, we have noticed that MLP and Random Forest models outperform SVM and KNN classifier for the considered dataset. The margin for tweaking of the KNN classifier's is less, hence performance is restricted, but in MLP classifier including tuning of hyper parameters, tweaking of weights in hidden layers with back propagation offer larger scope in reducing the error. It is also experimentally found that by comparing the results use of Feature set 2 (prepared by using SVD) can be considered better than that of statistical variance (feature set1). No overfit or

underfit observed in the model's performance except KNN when operated on variance (feature set1) exhibiting slightly underfit with a deviation of 4.4%. With the observation made on experimental results, it is concluded that MLP and Random Forest exhibit superior performance in speech digit utterance recognition and classification on both the feature sets. This work finds applications in HMI/Robotics, also to issue audio command in number format.

### Acknowledgement

Authors would like to thank RNS Institute of Technology for providing R&D facilities


### References

- [1] N.M.Nawi et al., "The Effect of Pre-Processing Techniques and Optimal Parameters selection on Back Propagation Neural Networks". International Journal on Advanced Science, Engineering and Information Technology, Vol.7 (2017) No. 3 ISSN: 2088-5334
- [2] Yamamaoto E, Nakamura S and Shikano K, "Lip movement synthesis from speech based on hidden Markov models," In proc. IEEE International Conference on Automatic Face and Gesture Recognition, pp.154-159, 1998.
- [3] G. S. V. S. Sivaram, Hynek Hermansky, "Sparse Multilayer Perceptron for Phoneme Recognition" IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 23-29, 2012.
- [4] M. S. Daud, I. M. Yassin, A. Zabidi, M. A. Johari, M. K. M. Salleh "Investigation of MFCC Feature Representation for Classification of Spoken Letters using Multi-Layer Perceptron (MLP)". International Conference on Computer Applications and Industrial Electronics (ICCAIE 2011) 978-1-4577-2059-8/11 ©2011 IEEE.
- [5] Jihyuck Jo, Hoyoung Yoo, and In-Cheol Park. "Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems" IEEE transactions on very large scale integration (VLSI) systems, 1063-8210 © 2015 IEEE.
- [6] Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K.J, "Phoneme Recognition Using Time-Delay Neural Network" IEEE Transactions on Acoustic, Speech, and signal processing, Vol.37, No.3, March 1989.
- [7] Sayf A. Majeed et al., "Mel frequency cepstral coefficients (mfcc) feature extraction enhancement in the application of speech recognition a comparison study" Journal of theoretical and Applied Information Technology 10th September 2015. Vol.79. No.1 © 2005 – 2015, ISSN:1992-8645.
- [8] Mahesh Goyani et. al "Performance Enhancement in Lip Synchronization Using MFCC Parameters", International Journal of Engineering Science and Technology, Vol.2(6), 2010,2364-2369
- [9] S. B. Davis and P. Mermelstein, "Comparision of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing. 28, no. 4, 1980.
- [10] Srikanth G N., M K Venkatesha, 'SEOA DRN: Social Exponential Optimization Algorithm Based Deep Residual Network for Visual Speech Recognition'. SSRG International Journal of Electrical and Electronics Engineering (SSRG-IJEEE), ISSN: 2348-8379. Volume 10 Issue 1, 90-105, January 2023
- [11] Nitin Trivedi et al. "Speech Recognition by Wavelet Analysis" International Journal of Computer Applications (0975 – 8887) Vol 15– No.8, February 2011
- [12] Gholamreza Farahani, "Robust Feature Extraction using Autocorrelation Domain for Noisy Speech Recognition". Signal & Image Processing an International Journal (SIPIJ) Vol.8, No.1, February 2017.
- [13] K.Pavan Raju1 , A. Sri Krishna2 And M. Murali3. "Automatic Speech Recognition System Using MFCC-Based LPC Approach with Back Propagated Artificial Neural Networks". ICTACT Journal On Soft Computing, July 2020, Vol. 10, Issue: 04 , ISSN: 2229-6956.

- [14] Shivam Sharma” Speech Recognition with Hidden Markov Model: A Review” International Journal of Scientific & Engineering Research, Vol..6, Issue 11, November-2015, ISSN 2229-5518
- [15] Datta Rakshith K. S, Rudresh M. D and Shashibhushan G, “Comparative performance analysis for speech digit recognition based on MFCC and vector quantization,” Global Transitions Proceedings, vol. 2, pp. 513–519, 2021.
- [16] Rathinavelu Chengalvarayan, Member, IEEE, and Li Deng, Senior Member, IEEE. “HMM-Based Speech Recognition Using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features”. IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 3, May 1997.
- [17] Chandralika Chakraborty, P.H. Talukdar “Issues and Limitations of HMM in Speech Processing: A Survey,” International Journal of Computer Applications (0975 – 8887) Volume 141 – No.7, May 2016
- [18] K. F. Lee and H. W. Hon, “Speaker-independent phone recognition using hidden Markov models,” IEEE Trans. Acoust., Speech, Signal Process., vol.37, No. 11, pp. 1641–1648, Nov. 1989.
- [19] Li Deng, Fellow, IEEE, and Xiao Li, Member, IEEE. “Machine Learning Paradigms for Speech Recognition: An Overview”. IEEE Transactions on Audio, Speech, And Language Processing, Vol. 21, No. 5, May 2013
- [20] Rizwan, Muhammad; Anderson, David V. “Using k-Nearest Neighbor and Speaker Ranking for Phoneme Prediction”. IEEE 13th International Conference on Machine Learning and Applications (ICMLA) - Detroit, MI, USA (2014.12.3-2014.12.6)] 2014.
- [21] Kumar, C., Ur Rehman, F., Kumar, S., Mehmood, A. and Shabir, G., “Analysis of MFCC and BFCC in a speaker identification system”, In proceedings 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-5, 2018.
- [22] Faragallah, O.S., “Robust noise MKMFCC–SVM automatic speaker identification”, International Journal of Speech Technology, vol.21, no.2, pp.185-192, 2018.
- [23] Taabish Gulzar, Anand Singh, Sandeep Sharma, “Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks”, International Journal of Computer Applications (0975 – 8887) Vol.101 No.12, September 2014
- [24] Khadar Nawas K, Manish Kumar Barik, Nayeemulla Khan. “Speaker Recognition using Random Forest”, ITM Web of Conferences 37, 01022 (2021)
- [25] Thambi, Sincy V.; Sreekumar, K. T.; Kumar, C. Santhosh; Raj, P. C Reghu. “Random forest algorithm for improving the performance of speech/non-speech detection”. First International Conference on Computational Systems and Communications (ICCSC)17-18 December 2014.


### Biographies of Authors



**Srikanth G N** , Received Engineering degree in Instrumentation Technology from SIT, Tumkur from Bangalore university in 1994 and MTech degree from SDMCET Dharwad from Visvesvaraya Technological University. currently pursuing PhD in RNSIT Research center under VTU. currently working as an Associate professor in Electronics and Instrumentation Engineering department, has 24 years of experience out of which 3 and half years of R & D experience in Biomedical industry, involved in the design and development of human blood parameter analyser. Teaching experience of 20 years, his research interests are in the area of Speech processing and image processing, control system. Published papers in the field of signal processing, speech and image processing. Instrumentation.

He can be contacted at email: srikanthgn27@gmail.com, srikanth.gn@rnsit.ac.in



**Dr. M K Venkatesha** , Director (since 2023), RNSIT, is a "JC Bose Gold Medallist" and has more than 42 years of teaching experience. He received Engineering degree from Mysore university in 1981, Post Graduation from the University of Manitoba Winnipeg, CANADA on a fellowship in 1986. Obtained Ph.D from University of Mysore in 1999. His research interests are in the area of signal processing and has published a number of research papers with many citations. Worked as a Registrar (Evaluation) of VTU, Belagavi (2001-04), Adviser of AICTE, New Delhi (2004-05) & Principal of BMSCE, Bengaluru (2005-08), Principal of RNSIT Bengaluru (2008-23).

He can be contacted at email: [mkvenkatesha@gmail.com](mailto:mkvenkatesha@gmail.com), [director@rnsit.ac.in](mailto:director@rnsit.ac.in)