Journal of Computational Science and Intelligent Technologies
eISSN: 2582-9041
DOI: 10.53409/MNAA/JCSIT/e202203031429
Type: Research Article

Performance Analysis of AI Models for Audio Digit Utterance Detection

Srikanth G N[1*], M K Venkatesha[2]

[1]Associate Professor, Department of Electronics & Instrumentation Engineering, R N S Institute of Technology, (Affiliated to Visvesvaraya Technological University), Bangalore, Karnataka, India.

[2]Principal, R N S Institute of Technology, (Affiliated to Visvesvaraya Technological University), Bangalore, Karnataka, India.

[*]Corresponding Author: Srikanth G N Email: srikanthgn27@gmail.com

Received: 12 June 2022; Accepted: 11 october 2022

Abstract: Automatic speech recognition has become integral to many applications, specifically HMI. Several AI models, supervised/unsupervised, are available in various platforms – MATLAB, LabVIEW, and Python with varying performance metrics. This paper presents a generic AI Model with the following features. The model is specifically built for 'digit word utterance detection.' The dataset contains an isolated set of digit utterances from 1 to 10 across speakers of different age groups. The dataset is prepared using established pre-processing steps, trimming each utterance to remove silence in the initial part and end of the utterance using Audacity software and noise removal. Also, single channel selection followed by sampling the speech signal at fs = 48kHz. Statistical variance analysis on the MFCC matrix for each digit utterance is carried out to obtain the feature set. KNN-based and MLP-based AI models are developed for this feature set (resulting from statistical methods). The performance of the developed AI models is analyzed. To reduce the computational complexity of the feature sets, dimensionality reduction has been applied to the extracted MFCC features using the SVD technique. This reduced number of principal components forms a new feature for the same utterances. KNN-based and MLP-based AI models are developed for this new feature set (resulting from the SVD method). The performance analysis is carried out for these models also. Results show that SVD with MLP performs better in classifying the uttered digit.

Keywords: MFCC, PCA, SVD, MLP, KNN, LPC, HMM

1 Introduction

Speech is considered a dominant mode of communication. It contains more information than text since it enables us to express internal feelings and emotions and convey the information. Automated speech recognition has been investigated over many years, aimed principally at realizing machines/Robots capable of recognizing speech and executing the instruction or commands given. Through continuous research in this field of advanced ASR systems, there is still a need and ample scope to improvise speech recognition engines. Speech waveforms are generally too complicated to compare and are considered non-stationary; however, when speech is investigated for shorter duration segments, it is assumed to be a stationary signal, meaning that statistical features of speech remain constant and finite. Therefore, short-time spectral density is usually extracted at short intervals and for analysis.

The techniques in recognition of speech involve pre-processing the speech signal by filtering, sampling, and applying transformations such that redundant and unwanted parts of the signal are suppressed or removed while preserving the information content. Although many methods are available for speech feature extraction, the most widely used method is to extract the MFCC (Mel Frequency Cepstral Coefficients) [1]-[4], which is a state of the art since it mimics the human perception of the speech. However, to track the dynamics of the speech, one can include the velocity and acceleration coefficients. Further, to obtain the speech's significant features and reduce the computation complexity, we need to apply dimensionality reduction techniques such as covariance on the transformation matrix, PCA, or the use of Singular value decomposition (SVD) on the MFCC matrix. The processed acoustic features are labeled and fed as an input to the machine learning algorithm with known output. We have applied supervised machine learning [4]. Some of the algorithms for the classification of speech recognition include Logistic regression and Multilayer perceptron (MLP); however, DNN demands a huge dataset for training to achieve respectable classification and prediction. For the multiclass problem, as in our work, concerning the database's size, the most suitable classification technique is MLP.

MLP is a feed-forward artificial neural network (ANN) class where every node uses a nonlinear activation function. It utilizes a supervised feed-forward network with backpropagation for training the network by updating the weights to minimize the loss function. In this paper, we have considered two feature sets for the entire digit utterance dataset generated from MFCC, i.e., SVD features and variance of the chosen variables from the variance-covariance matrix of MFCC. Experimental analysis was conducted using KNN and MLP classifiers. The performance of MLP is better than KNN in classifying digit utterances. The influence of the work can be seen in applications such as Human-machine communication, Robotics, Advanced speech recognition engines, etc. The rest of this paper is presented as follows. Related work is summarized in Section 2 and ends with framed Objectives. Section 3 presents the proposed modeling, including feature extraction and a description of the developed models. Section 4 is about the Experimental results achieved by the chosen classifiers. Finally, discussions and conclusions are provided in Section 5.

2 Related Work

The ASR systems aimed at recognizing the phonemes, words, and hence sentence decoding for speaker identification adopting various algorithms hitherto LPC [8] with HMM [7][9] wherein predicting the output based on expectation maximization thereby reducing the error. To extract the speech features, many researchers used MFCC [1][2][3][4][8][10] and found it to be the most suitable way to extract them. In [19], it is shown that MFCC outperforms LPC. Further, machine learning approaches like MLP and DNN are primarily based on many data features for ASR.

In ASR systems, hitherto, researchers have used Hidden Markov Model in predicting speech utterances, where phoneme is considered a basic unit of speech and its combinations form the word utterance. In HMM, the states are generated for the word following the transition probabilities during the training phase. In the testing cycle for a given input, the probability that the target sequence is generated from each vocabulary word is calculated, and the identification is done based on the highest accumulated probability value. However, it is noticed that the Hidden Markov model (HMM) uses a Gaussian mixture output distribution, producing diminishing returns and accuracy [13], especially because the training phase is complicated and computationally intensive. Any misalignment of the states may lead to a fault in the recognition phase [6][7][9]. It is ascertained that the use of machine learning algorithms as a substitute to HMM will improvise the ASR systems and maintains promise to enhance ASR generation substantively by rigorously checking the effectiveness of the advanced strategies [1][3][4]. Machine learning algorithms are successfully applied in classifying uttered speech signals about speech detection and speaker identification systems with a considerable accuracy ranging from 65 to 98%, depending on the language used, dataset size, features used in a training phase, and applied ML algorithms. The methods applied for speech processing are well narrated in [14].

For isolated word recognition, apart from applying DCT, wavelet transform is also used for feature extraction, reducing the complexity of neural network calculation by minimizing the word utterance into

lower dimension features vector [20].

The main objectives of the work are as follows.

i. To extract the robust significant features considering statistical variations about the data derived from speech utterance.
ii. Classify digit utterances using machine learning algorithms aimed at higher accuracy.

3 PROPOSED MODELLING

Most audio signal processing and recognition systems incorporate important and robust feature selection and extraction as a major module. In the feature extraction process, short segments of speech usually (30 – 40 msec) are extracted in succession from word utterance, which is utilized to depict the audio signal. In the proposed method, speech features are derived using the MFCC model.

The purpose of the system here is to convert speech waveform into parametric form for further investigation. As mentioned earlier, the short segments of the speech are considered in succession for analysis since it is assumed that speech is statistically stationary in this small duration. We also get reliable spectral data due to a limited number of samples. However, for a longer duration, speech properties will change. Hence Short Time Fourier Transform (STFT) is considered for analyzing the speech in the spectral domain. The window that is generally used is the hamming window since it has good side lobe attenuation. Window overlapping width of 25% accounts for 10 msec, provides a smooth transition, and avoids spectral distortion. Windowing operation is expressed in (1). If w(n) represents a window defined for duration $0 \leq n \leq N \leq 1$, where N is the total number of samples in each segment, then the result of the windowing process is

$$y(n) = x(n) \,.\, w(n) \qquad 0 \leq n \leq N\text{-}1 \tag{1}$$

Hamming window made of cosine term is shown in (2).

$$w(n) = 0.54 - 0.46\, cos\left[\frac{2n\pi}{N-1}\right] \quad 0 \leq n \leq N\text{-}1 \tag{2}$$

Here, the N represents the total number of samples in the block. Various methods are available to represent the speech signal parametrically. The ASR systems and speaker identification applications include autocorrelation analysis and LPC analysis [2][8][18]. It is shown that features extracted using MFCC [8] perform better than LPC [19] in speech recognition.

We have used Mel frequency cepstral coefficients (MFCC) to extract features from the speech signal. The variance-covariance matrix is calculated from the MFCC matrix, and the chosen variables' statistical variance is extracted to form an acoustic feature set to train the machine.

Apart from Variance features to form a new feature set, we have also executed Singular value decomposition (SVD) on the overextended MFCC matrix, and most significant singular values are retained to form another set of acoustic features related to information content and unwanted redundant values are removed which are included due to noise. Word recognition and classification are implemented using KNN and MLP on the mentioned feature sets. The steps implemented for feature extraction, recognition, and classification of the digit utterance are shown in Fig.1, and the algorithm steps followed are depicted in table 1 & 2

3.1 mfcc

The block diagram shown in Fig.1 depicts steps to obtain MFCC coefficients for speech utterance. The human speech signal is known to be within 4kHz, which includes most of the human-generated sound energy. Hence a typical sampling rate of 8kHz will suffice. However, in our experiments sampling rate of 48 kHz is considered since it is compatible with most of today's systems with modern machine interfaces. To derive the information content from the speech signal, we have chosen MFCC since MFCC mimics the functional characteristics of the human ear and sound perception, i.e., it has a linear response up to 1000 Hz and turns to be logarithmic after 1kHz.

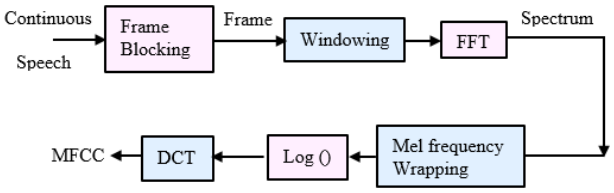

Figure 1: Steps followed to obtain MFCC

As mentioned, we use short speech segments out of complete utterances in succession and process each segment to get reliable features. FFT is applied on windowed signals to transform each segment of N samples to the frequency domain for investigating spectral content. Studies and experiments show that human speech perception in the frequency domain is nonlinear after 1kHz. Therefore, for each tone signal with a true frequency of f, the subjective pitch or frequency is determined using the "Mel" scale. The equation for frequency mapping from linear scale to Mel scale is expressed in (3).

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

The inverse equation is given by (4) to get the original frequency.

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

The Mel scale is a linear frequency scale below 1 kHz and a logarithmic scale above 1 kHz. We use a filter bank, a series of triangular bandpass filters to mimic Mel's spectrum. Each filter and center frequency in the filter bank are evenly spaced on the mel scale. The distance between the center frequency and the bandwidth determines the constant spacing of Mel frequencies. To calculate the strength of the filter bank, we multiply each filter bank with a power spectrum and then add the coefficient. A set of MFCCs is obtained by applying DCT. A DCT is required to decorrelate the filter bank coefficients. Denoting the Mel power spectral coefficients resulting from the last step as

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

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

For each frame, the frame audio feature vector contains 13 real scalars chosen as above, typically including the delta coefficients. For better resolution and delta-delta coefficients to identify the dynamics involved in the speech signal, we have experimentally determined that only by using cepstral coefficients, we have achieved good results, favoring reduced dimension of the matrix; hence velocity and acceleration coefficients are not included.

3.2 Feature extraction

From the generated MFCC coefficients, we have chosen the first 13 coefficients, excluding the first energy element since it provides information about the RMS energy rather than the vocal-tract configuration hence the utterance is transformed to a sequence of acoustic vectors forming a matrix with 13 columns and number of rows populated depending on the length of utterance, i.e., each utterance may have a varying time duration, each vector is representing information in the small time window of the signal. The regular/native methods involve applying the machine learning algorithm (supervised or unsupervised) on this complete matrix with varying rows, e.g., Digit 8 utterance has 19X13 and digit nine have 21X13, the difference in time related to the difference in the number of rows.

In the proposed method, these over-extended MFCC matrices are processed in two techniques to extract the significant feature sets.

i) We have also considered the statistical variations of the chosen variables in the proposed algorithm to form a fixed-length vector. The steps followed are, Obtain the covariance matrix from the MFCC matrix formed by considering an average of every 3 rows set from the overextended MFCC, and the variance-covariance matrix is calculated from the obtained from the MFCC matrix. The general representation of covariance considering two random variables is expressed in equation (6).

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

The next step is to Extract the diagonal elements of the covariance matrix, giving the variances of the selected variables associated with the utterance. These post-processed vectors are referred to as input vectors and populated in a .csv file labeling the input variables of the formed acoustic vector and the known output. Table 1 shows the steps followed in digit utterance recognition and classification. Note: The above steps are similar to PCA but not exactly PCA.

In the proposed method by experimentation, we observed that the 1st nine variables (C1 to C9) give more than 90% of the variance, so irrespective of the size of the uttered digits in the proposed algorithm, dimensionality reduction is achieved by choosing the main diagonal elements of the MFCC covariance matrix which is different from the presented literature on PCA[21] were in the eigenvalues based on the input features/parameters are obtained and based on the threshold the number of principal components is chosen. In PCA, eigenvalues of each row of DCT output are taken, but in the proposed algorithm, we are taking MFCC to correspond to each row (a row corresponds to an overlapping segment of the input signal). Hence, the computation time is exponentially reduced, and distinct signature vectors are obtained for each digit. Also, the difference is observed for different digit utterances, as depicted in Fig 2 to 4. This leads to improved performance metrics over the existing methods, demonstrated in the next section, where in KNN, MLP algorithms are applied, and the corresponding results are shown in Table 5.

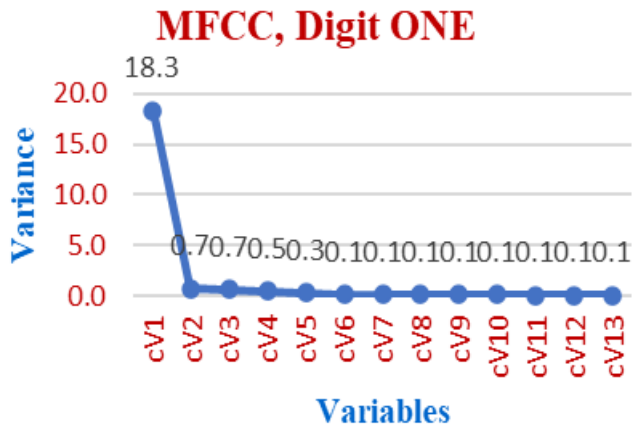

Figure 2: Variance values from MFCC, Covariance matrix

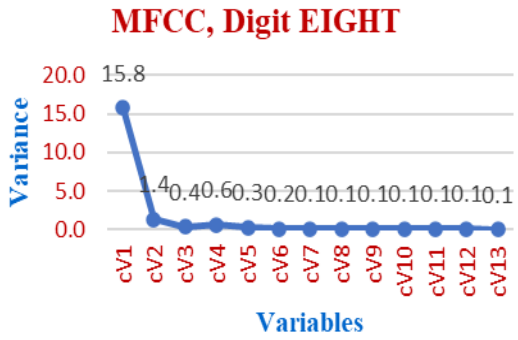

Figure 3: Variance values from MFCC, Covariance matrix

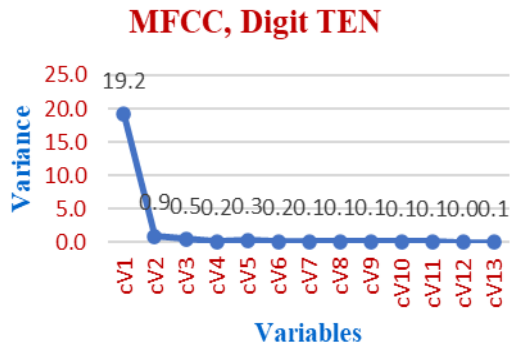

Figure 4: Variance values from MFCC, Covariance matrix

Table 1: Steps followed for Digit utterance recognition using MLP and KNN on the statistical Variance feature set

1.	**Input:** Digit Utterance Audio .mp4 file
2.	Sampling audio signal with fs=48kHz
3.	Windowing using Hamming window w(n)
4.	Generate MFCC coefficients matrix
5.	Generate variance-covariance matrix on MFCC
6.	Extraction of Variance C1 to C13, populate .csv file
7a.	Apply MLP on features from step 6, with 3 hidden layers with activation function tanh, tuning of hyperparameters.
7b.	Obtain the performance metrics of the AI model
8.	Repeat step 7 with the KNN algorithm
9.	**Output:** Classification Result Evaluation by metrics and Algorithm performance comparison.

ii) As mentioned earlier, to reduce the computation complexity involved in obtaining the significant robust features from speech, we have applied singular value decomposition on an over-extended MFCC data matrix. SVD will decompose the input matrix into a set of orthogonal matrices and a diagonal matrix, and the equation is shown in (7)

$$A = U\ S\ V^T \tag{7}$$

Here, 'A' is the MFCC matrix.

'U' is the orthonormal matrix (m x n matrix of the orthonormal eigenvectors of AA^T).

'V' is the orthonormal matrix (transpose of a n x n matrix containing the orthogonal eigenvectors of A^TA.).

'S' is the diagonal matrix containing singular values arranged in descending order to form the features set. Singular values are the square root of the eigenvalues. SVD will decompose the data matrix (it need not be a square matrix) to a low-rank matrix. A further approximation is obtained by retaining a specific number of singular values, which is arranged in descending order ($\sigma_1 > \sigma_2 > \sigma_3 \ldots. \sigma_p$) and removing all the redundant values less than σp that are included due to the presence of noise. It is found that the feature set obtained from SVD analysis will elevate the word recognition accuracy.

We have considered the singular values as a data matrix to reduce the computation complexity and populated in the .csv file comprising labeled input and output for the entire dataset chosen.

The samples of the signature vectors obtained from SVD analysis with the singular values are shown for digit utterances 1,5,8 in Fig 5 to 7.

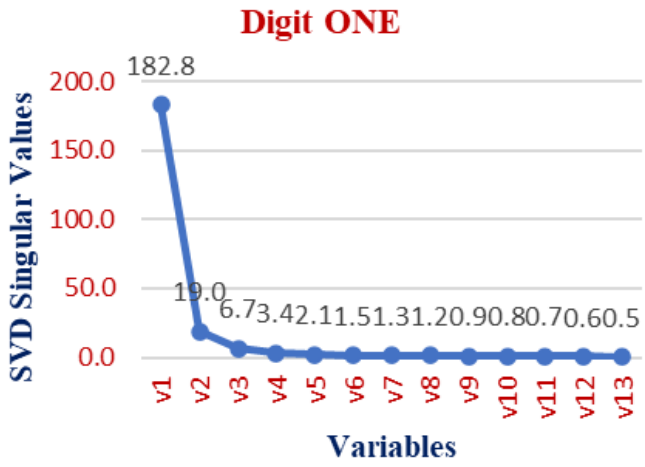

Figure 5: Singular values from SVD analysis on MFCC matrix

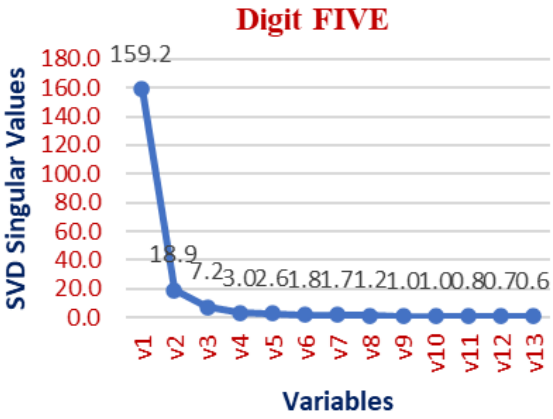

Figure 6: Singular values from SVD analysis on MFCC matrix

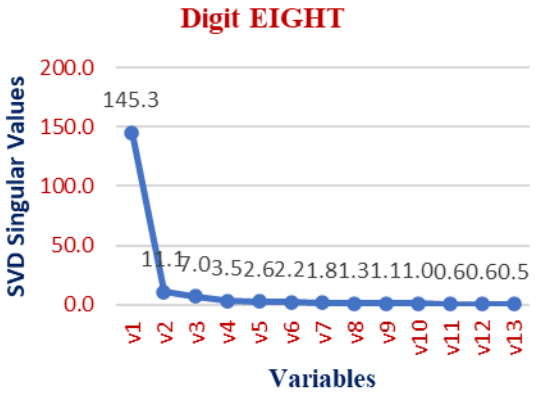

Figure 7: Singular values from SVD analysis on MFCC matrix

Table 2: Steps followed for Digit utterance recognition using MLP and KNN on the SVD feature set

1.	**Input:** Digit Utterance Audio .mp4 file
2.	Sampling audio signal with fs=48kHz
3.	Windowing using Hamming window w(n)
4.	Generate MFCC coefficients matrix
5.	Obtain SVD on over-extended MFCC matrix
6.	Extract Singular values of variables in descending order and populate the .csv file
7a.	Apply MLP on features from step 6, with 3 hidden layers with activation function tanh, tuning of hyperparameters.
7b.	Obtain the performance metrics of the AI model
8.	Repeat step 7 with the KNN algorithm
9.	Output: Compare the performance of the two models.

The obtained feature set is used as an input to the KNN and MLP algorithms, and the respective results are shown in Table 5. The algorithms are implemented in Python to classify and predict the digit utterance. Table 2 shows the steps followed in digit utterance recognition and classification.

3.3 knn classifier

K-Nearest Neighbor is a supervised machine-learning class. The data to be predicted (k = 5 in this experiment) are assigned to a cluster of centroids based on the similarity measure using the distance function. KNN provides highly adaptive behavior and is optimal in the large sample limit. The downside is that it is computationally intensive and requires a lot of storage. KNN is a traditional classification method and does not require any training effort, it strongly depends on the quality of measurements between samples, and noise can easily affect the performance of this classifier. The distance function used here is the most common matrix Euclidean distance, shown in equation (8), which is useful in a low-dimensional dataset.

$$d_{\text{euclidean}}(x, y) = \sqrt{\sum_{i} (x_i - y_i)^2} \tag{8}$$

where x and y are m dimensional vectors and denoted by $x = (x_1, x_2, x_3 \ldots x_m)$ and $y = (y_1, y_2, y_3 \ldots y_m)$ represents m attributes of two records. We have also used the KNN classifier for recognizing digit utterance based on the feature sets developed using statistical variance, and Table 5 and Fig. 9 and Fig.11 depict the results.

3.4 multilayer perceptron (MLP) classifier

The multilayer perceptron forms a network consisting of an input layer, one or more hidden layers of computation nodes, and an output layer. It is a feed-forward artificial neural network (ANN) with a supervised learning strategy (having labeled input-output in tabular format or .csv file) and using backpropagation method network weights are tuned to minimize the loss function and to attain the convergence [16], hence driving the error towards zero. Multiple layers of MLP with tanh, a nonlinear activation, and ReLU distinguish MLP from linear perceptron.

By experimentation, we found that promising results are given by rectified linear activation function. Relu is used in hidden layers of MLP. ReLU overcomes the vanishing gradient problem in SGD and allows the models to learn faster and perform better. In addition, we have found that tanh as an activation function showed markable accuracy compared to the use of ReLU, probably since the derivatives are not monotonic, tanh solve the problem of dead neuron, which is a requirement in any neural network using backpropagation to minimize the error, results shown are an output of MLP algorithm by using the tanh activation function. The above implementation is based on the limited dataset consisting of digit utterances from 1 to 10 from 36 subjects forming 345 .mp4 files.

MATLAB is used to develop a program for speech processing, MFCC coefficient generation, and extraction of statistical variance to form a significant feature set concerning digit utterances, same is entered in .csv files.

SVD analysis is performed based on the overextended MFCC matrix for the whole dataset, and the principal values are populated in the .csv file forming 2nd feature set.

The mentioned feature sets are used with a 75%:20% ratio for training and testing on MLP and KNN algorithms separately. Table 3 and 5 and Fig. 8 and Fig. 10 depicts the results.

The input vectors train the MLP classifier consisting of three hidden layers and tanh as an activation function. The program for digit classification and prediction is developed using Python. Through experimentation considering the variance of the chosen variables C1 to C13, we obtained an accuracy of prediction 77%, dropping C10 - C13 and considering the most significant MFCC coefficients C1 to C9 MLP classifier a learning rate equal to 0.001. For 500 iterations, convergence is obtained with improved accuracy. Using a second feature set developed with SVD analysis on the MLP classifier resulted in an

accuracy of 99%. Table 5 gives the detailed result analysis considering the prepared different feature sets and the applied algorithms.

3.5 dataset

The proposed model uses a speech digit utterance dataset. The data collection consists of 345 audio files in .mp4 format formed from 36 subjects, including males and females of different age groups (excluding children), uttering ten sentences. The sentence includes digit utterances from one to ten, forming 345 files. From these sentences' word utterances are extracted using 'Audacity,' an audio-processing software tool.

4 RESULTS AND DISCUSSIONS

In this work, we obtained MFCC coefficients for the entire digit utterance dataset. Fig 2 to Fig 4 display the samples of the MFCC-generated coefficients for the digit utterances one, eight, and ten.

We have implemented MLP and KNN algorithms to predict and classify the feature sets obtained using SVD and statistical variance. The confusion matrix in Table 3 exhibits digit recognition against the test samples using an MLP classifier based on the SVD feature set. The principle diagonal elements depict the degree of predicted output against the actual class. A markable classification level is observed in the testing cases except for one misclassification for digit 9, highlighting the performance of MLP.

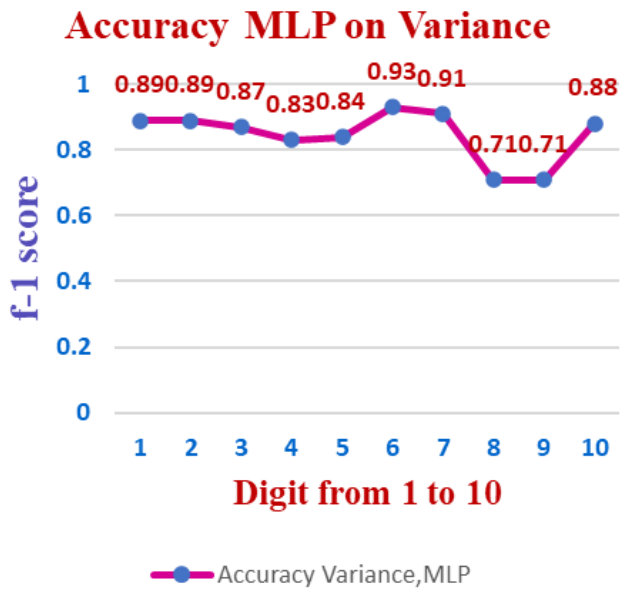

Figure 8: Recognition accuracy against uttered digits using Variance and MLP Classifier

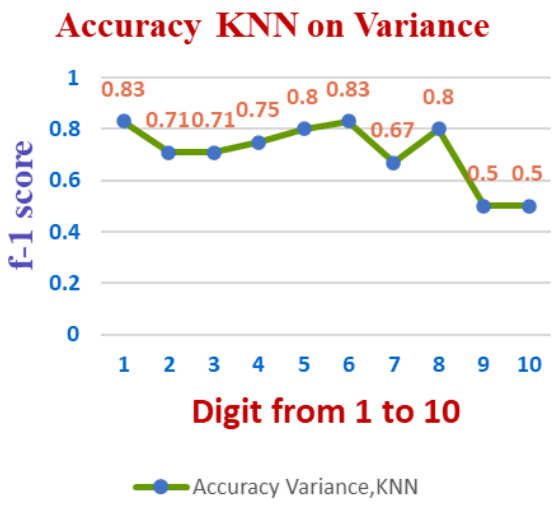

Figure 9: Recognition accuracy against uttered digits
Using Variance and KNN Classifier

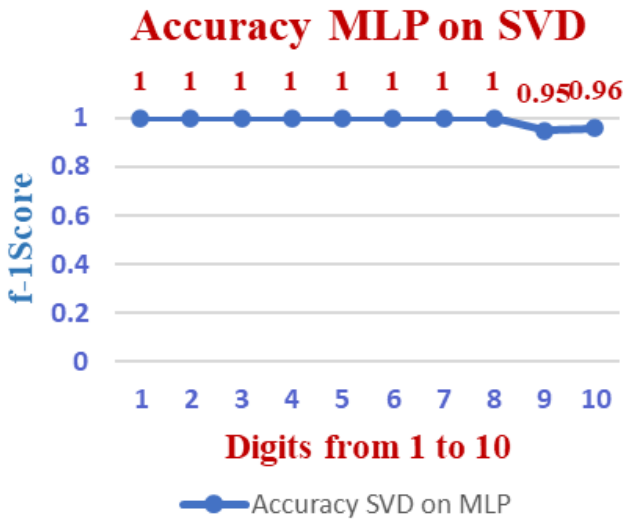

Figure 10: Recognition accuracy against uttered digits
Using SVD and MLP Classifier

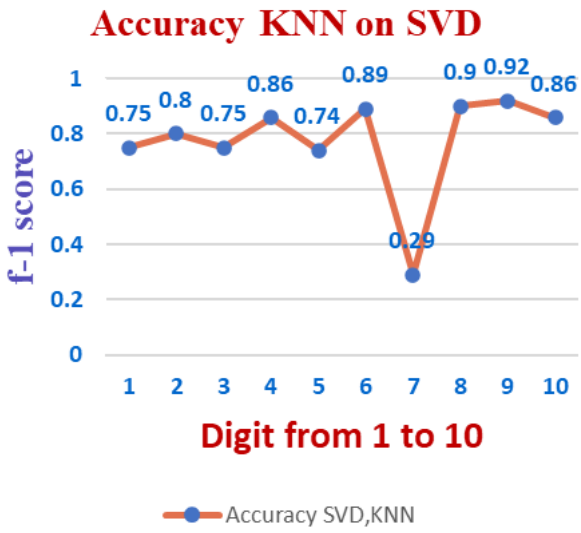

Figure 11: Recognition accuracy against uttered digits Using SVD and KNN Classifier

Table 3 Confusion Matrix of Digit Classification Using MLP based on the SVD feature set

		PREDICTED BY MODEL									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ACTUAL CLASS	**1**	**11**	0	0	0	0	0	0	0	0	0
	2	0	**9**	0	0	0	0	0	0	0	0
	3	0	0	**16**	0	0	0	0	0	0	0
	4	0	0	0	**8**	0	0	0	0	0	0
	5	0	0	0	0	**12**	0	0	0	0	0
	6	0	0	0	0	0	**10**	0	0	0	0
	7	0	0	0	0	0	0	**12**	0	0	0
	8	0	0	0	0	0	0	0	**14**	0	0
	9	0	0	0	0	0	0	0	0	**10**	1
	10	0	0	0	0	0	0	0	0	0	**12**

Table 4 shows the detailed classification report wherein a respectable F1-score is obtained. Figures 8 to 11 display the accuracy of AI algorithms and the chosen feature sets. Fig.12 portrays the overall performance of the Algorithms.

Table 5 compares the accuracy attained by algorithms on the feature set obtained using SVD and statistical variance. Using MLP, the accuracy obtained in digit utterance recognition is 99%. With the same features when the KNN classifier is used, we observed the degradation in recognition accuracy by

around 12 % compared to the MLP classifier highlighting the merits of MLP. The performance of KNN depends on the quality of the samples, i.e., little noise included can influence the performance further, each attribute is treated as totally different from all of the attributes, w.r.t [8] our model shows an increase in prediction accuracy of 10% by using MFCC, and w.r.t LPC analysis 22%.

Table 4: Classification-Report for Digit Utterance Recognition Using MLP.

Metric / Class	Precision	Recall	f1-score	Support
1	1.00	1.00	1.00	11
2	1.00	1.00	1.00	9
3	1.00	1.00	1.00	16
4	1.00	1.00	1.00	8
5	1.00	1.00	1.00	12
6	1.00	1.00	1.00	10
7	1.00	1.00	1.00	12
8	1.00	1.00	1.00	14
9	1.00	0.91	0.95	11
10	0.92	1.00	0.96	12
accuracy			**0.99**	115
macro avg	0.99	0.99	0.99	115
weighted avg	0.99	0.99	0.99	115

Mean squared error: **0.0086**

Table 5: Performance of Digit utterance classifiers systems based on SVD and Statistical Variance from MFCC

Digit Data	Feature set on MFCC	Classifier	Recog Rate %	MSE
1 – 10	SVD	MLP	99	0.008
1 – 10	SVD	KNN	87	0.456
1 – 10	Variance	MLP	85	0.28
1 – 10	Variance	KNN	75.00	0.33

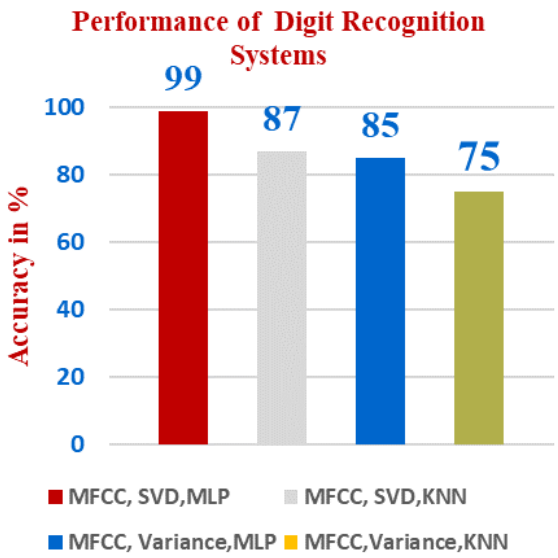

Figure 12: Performance of algorithms with Features from SVD and MFCC statistical variance

5 CONCLUSION

The Experimental results justify that the utterance of the digits was recognized and classified with an accuracy of 99%. It is noticed that MLP outperforms the KNN classifier for the database considered. The scope for fine-tuning the KNN classifier's performance is much limited. In comparison, the adjustable number of hidden layer neurons and tunable hyperparameters of MLP with backpropagation provide a wider opportunity to decide for classification by minimizing the loss function. However, the system must work effectively even when speech is recorded under a noisy environment or corrupted by noise when it is received, therefore augmenting lip movement, meaning that the geometrical/transform features obtained from video frames pertained to ROI (lip portion) along with corresponding audio features for the word utterances is expected to improvise the performance of speech recognition systems.

Acknowledgments: We are very thankful to RNS Institute of Technology's management for providing us with the R&D facilities.

Funding Statement: The author(s) received no specific funding for this study.

Availability of Data and Materials: The data used to support the findings of this study can be obtained from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Jo, H. Yoo and I. -C. Park, "Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 2, pp. 754-758, 2016. https://doi.org/10.1109/TVLSI.2015.2413454

[2] M. M. Goyani, N. M. Patel, M. A. Zaveri, and S. Vallabhbhai, "Performance Enhancement in Lip Synchronization Using MFCC Parameters," *International Journal of Engineering Science and Technology*, vol.2, no. 6, pp. 2364-2369, 2010. https://bit.ly/3I4aMSm

[3] S. A. Majeed, H. Husain, S. A. Samad, T. F. Idbeaa, "Mel frequency cepstral coefficients (mfcc) feature extraction enhancement in the application of speech recognition a comparison study," *Journal of theoretical and Applied Information Technology*, vol. 79, no. 1, pp.38-56, 2015. https://bit.ly/3YUF28D

[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)," *In. IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE 2011)*, Penang, Malaysia, pp. 16-20, 2011. https://doi.org/10.1109/ICCAIE.2011.6162096

[5] 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. https://doi.org/10.1109/TASL.2011.2129510

[6] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 37, no. 11, pp. 1641–1648, 1989. https://doi.org/10.1109/ICASSP.1993.319355

[7] S. Sharma, "Speech Recognition with Hidden Markov Model: A Review," *International Journal of Scientific and Engineering Research*, vol. 6, no. 11, pp. 185-190, 2015. https://bit.ly/3PN19t9

[8] K. P. Raju, A. S. Krishna, and M. Murali. "Automatic Speech Recognition System Using MFCC-Based LPC Approach with Back Propagated Artificial Neural Networks," *ICTACT Journal on Soft Computing*, vol. 10, no. 04, pp. 2153-2159, 2020. https://doi.org/10.21917/ijsc.2020.0306

[9] R. Chengalvarayan and L. Deng, "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, pp. 243–256, 1997. https://doi.org/10.1109/89.568731

[10] J. Pinto, B. Yegnanarayana, H. Hermansky and M. M. Doss, "Exploiting contextual information for improved phoneme recognition," *In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, pp. 4449-4452, 2008. https://doi.org/10.1109/ICASSP.2008.4518643

[11] J. Pinto, G. S. V. S. Sivaram and M. M. Doss, H. Hermansky and H. Bourlard, "Analyzing MLP based hierarchical phoneme posterior probability estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, No. 2, pp. 225–241, 2011. https://doi.org/10.1109/TASL.2010.2045943

[12] H. Ketabdar and H. Bourlard, "Enhanced phone posteriors for improving speech recognition systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.18, no.6, pp.1094–1106, 2010. https://doi.org/10.1109/TASL.2009.2023162

[13] X. Huang and L. Deng, "An overview of modern speech recognition," *Handbook of Natural Language Processing*, Second Edition. FL, USA: Microsoft Corporation, Chapter No. 15, Section No. 1, Page No. 339-366, 2009. https://bit.ly/3PNdKwC

[14] L. Deng and X. Li, "Machine Learning Paradigms for Speech Recognition: An Overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060-1089, 2013. https://doi.org/10.1109/TASL.2013.2244083

[15] M. Rizwan and D. V. Anderson, "Using k-Nearest Neighbor and Speaker Ranking for Phoneme Prediction". *In IEEE 13th International Conference on Machine Learning and Applications (ICMLA)*, Detroit, MI, USA, pp. 383-387, 2014. https://doi.org/10.1109/ICMLA.2014.68

[16] N. M. Nawi, A. S. Hussein, N. A. Samsudin, N. A. Hamid, M. A. M. Yunus *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, no. 3, pp. 770-777, 2017. https://doi.org/10.18517/IJASEIT.7.3.2074

[17] A. F. Kuri-Morales, "The Best Neural Network Architecture," *13th Mexican International Conference on Artificial Intelligence, MICAI2014*, Tuxtla Gutiérrez, Mexico, pp. 72-84, 2014. https://doi.org/10.1007/978-3-319-13650-9_7

[18] G. Farahani, "Robust Feature Extraction using Autocorrelation Domain for Noisy Speech Recognition," *Signal and Image Processing: An International Journal*, vol. 8, no.1, pp. 23-44, 2017. https://doi.org/10.5121/sipij.2017.8103

[19] 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*, vol. 28, no. 4, pp. 357–366, 1980. https://doi.org/10.1109/TASSP.1980.1163420

[20] N. Trivedi, V. Kumar, S. Singh, S. Ahuja and R. Chadha, "Speech Recognition by Wavelet Analysis," *International Journal of Computer Applications*, vol. 15, no. 8, pp. 27-32, 2011. http://dx.doi.org/10.5120/1968-2635

[21] A. Winursito, R. Hidayat and A. Bejo "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," *In. International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, pp. 379-383, 2018. https://doi.org/10.1109/ICOIACT.2018.8350748