# Performance of Isolated and Continuous Digit Recognition System using Kaldi Toolkit

**Mahadevaswamy, D J Ravi**

***Abstract:*** ***A digit recognition system is built for recognizing the sequence of digits through 0-9. The system is experimented with speech corpus created in the room environment. The acoustic information to feature representation is achieved using PLP and MFCC features. The system initially utilized the conventional GMM-HMM framework, state of the art hybrid classifier with varied number of states to complete the speech recognition task, i.e., the system is first trained and tested using Monophone models, and system's recognition accuracy is then evaluated using Triphone Models: Triphone1 models, which was later followed by Triphones2 models and Triphones3 Models. The N-gram Language model is used for both Monophone and Triphone training. The system performance is evaluated with the use of MFCC and PLP parameterisation techniques on Kaldi toolkit. The system performance is evaluated using metrics word error rate (WER) and Word Recognition Accuracy (WRA). The proposed system can be utilized for building speech applications.***

***Keywords: PLP, MFCC, GMM-HMM and KALDI TOOLKIT.***

## I. INTRODUCTION

Communication ability is a crucial activity for human beings. The physiological process of speech production, perception and natural speech are god's gift to the mankind for expressing thoughts of individual minds in an effective way. The natural speech and silence acts as chain links for connecting and disconnecting human mouth and ears respectively. Automated human speech recognition (ASR) is the technology for mapping spoken information into text form. The ASR has been quickly-growing and broadly spreading for several languages throughout the world. It is available in several languages namely, Kannada, Tamil, Telugu, Malayalam, English, Korean, Mandarin, Hindi, Spanish, Japanese, Persian, Arabic etc. The applications not requiring physical connection like speech recognition over mobile phones have impressed the speech processing community through miraculous performance. The ASR systems give tremendous performance in clean conditions. But, despite of significant improvements with respect to the technology, the performance of speech to text conversion algorithms is degraded by the uncontrolled noise, especially when speech sounds are considered from the high noise conditions.

**Revised Manuscript Received on June 22, 2019.**

**Mahadevaswamy,** Research Scholar, Dept. of Electronics and Communication,Vidyavardhaka College of Engineering, Mysuru, Visvesvaraya Technological University, Belagavi

**D J Ravi,** Dean Academics & Professor of ECE, Dept. of Electronics & Communication, Vidyavardhaka College of Engineering, Mysuru, Visvesvaraya Technological University, Belagavi

The speech community has reported wide variety of speech recognition systems for few Indian languages. The report of Indian census 2001 reveals the reduction in number of native speakers. The variants of ASR systems are unintegrated based on type of spoken information. The input may be continuously spoken signal or spontaneous signal. Many ASR products namely Alexa, Google's Voice Assistant, i-Phone's Siri, Microsoft's Cortana are publically available in English and Foreign languages but the goal is not reached in Indian languages and the many readily available ASR products such as Alexa and Google voice assistant still needs improvement. The performance of ASR systems degrades mainly due to recording medium, additive noise, dialect, native language influences, accent, age, low/high technical skill, health and emotional state of speakers. Speech recognition accuracy is a crucial feature in speech applications activated by voice.

## II. LITERATURE SURVEY

The acoustic models based on TDNN trained using criterion of LF-MMI outperform that of DNN in various types of ASR attempts. KLD regularization serve as efficient method for adaptation of DNN based acoustic models and is used for training Time delay DNN acoustic models. The ASR system achieved 7-29% WER on Mandarin speech recognition task. The combination of KLD, TDNN, LF-MMI acoustic models lead to remarkable success especially when in domain data is limited [1].

A FFE-CMLLR is explored for combined normalization of speaker and the speaking environment. MFCC features are processed by LDA for dimensionality reduction are further transformed by MLLT. The DNN-HMM acoustic modelling is used to achieve recognition. The F-FE-CMLLR speaker and environment normalization techniques leads to significant improvements in the recognition accuracy above SAT features when applied to DNN-HMMs. Iterative type of training works better compared to sequential type of training irrespective mentioned test conditions [2].

A spontaneous speaker independent ASR system is reported for Punjabi language. ASR system is built using Sphinx Toolkit and live system model is created using Java programming. The database consists of 6012 words and 1433 sentences. The WER and SER are 93.79% and 90.8% respectively [3].

The feed-forward network based deep neural network (DNNs) architectures are investigated explored to analyze the impact of parameters like size of the model, architecture, and training details, on WER. The experiments are carried

out on Switchboard benchmark corpus to compare convolution networks to standard DNNs. Large DNN models were also built using combined large speech corpus of 2100 hours [4].

The speaker independent ASR system is built using Mel frequency Cepstral coefficients (MFCC) and word level, syllable level Hidden Markov Model (HMM) acoustic models. The experiments were carried out on Arabic Digits database (ARADIGITS-2) consists of 2704 clean utterances by 112 speakers. The syllable level acoustic units outperform word level units by an average word accuracy rate of 0.44 and 0.58% for clean condition and multi condition training tasks respectively [5].

An extraordinary effort has been put to build Chhattisgarhi speech corpus consists of 100 isolated spoken words, 67 sentences and 478 speakers. The word recognition accuracy of 99.84% and 94.24% is achieved using ANN and SVM respectively. An accuracy of 81.25% is obtained for HMM based continuous speech recognition [6].

The ASR system is built using MFCC and PLP parameterization principles for the recognition of continuous speech samples of Punjabi Language. The performance of the system is presented for monophone model and triphone's i.e., Tri1, Tri2 and Tri3 models. It is observed that the Tri3 models outperform Tri2 models and Tri2 models lead to improvements over Tri1 models and all triphone models outperformed the monophone models and MFCC features worked well than PLP features. The Tri2 and Tri3 models using MFCC achieved a best WER of 21.8% and 21.2% [7].

A SI isolated word recognition system is developed using two fusion techniques. The database is bi-furcated into a clean set and several noisy sets, to ensure that the system is robust to noise in real time scenarios where one rarely get noise free environment. The 13-dimensional, 26-dimensional, 39-dimensional MFCC features are used for the experimentation. The Fusion 2 technique outperformed the GMM-HMM, VQ and I-Vectors and Fusion 1 technique in the clean conditions for 13dimensional MFCC features [8].

Language independent neural network architecture is built here. A common framework is built for representing phones of various languages. Transfer learning and joint learning techniques are explored to adapt the networks trained using English with limited Bangla language data. The approach shown impressive performance across other languages of Bangla family and with few south Indian languages [9].

Proposed a Multilingual phoneme recognition model with four Indian languages namely, Odia, Telugu, Kannada and Bengali. The model recognizes phonemes irrespective of the language. The transcription from International phonetic alphabets employed for clustering of similar phonemes from multiple languages. The MFCC amalgamated with GMM-HMM and DNN architectures are utilized for experimentation. The DNN models able to shine over HMMs. The tandem Multi-lingual phoneme recognition systems outperformed the baseline phoneme recognition systems. The performance of baseline systems BN_OD, KN_TE and KN_TE_BN_OD in terms of phone error rate is 32.0, 33.1 and 35.1 respectively and that of corresponding tandem Multi-lingual phone recognition systems is 30.6, 31.9 and 34.1 respectively. A significant reduction of 1.4, 1.2 and 1.0 is achieved [10].

Punjabi ASR system is designed for use in mobile phone applications with a framework consisting of four various acoustic modeling approaches- context depended deleted interpolation, context dependent tied and untied, context independent. The system performance is reported for MFCC amalgamated with 4, 16, 32 and 64 Gaussian Mixture Models using WRA, WER and memory requirement. The context dependent untied model outperformed all other models with least WER. CI models need lesser amount of memory. It is confirmed that recognition results at higher number of GMMs is better than lower number of GMMs [11].

A DNN architecture based regression model is designed to determine the amplitude, phase for faithful reconstruction of speech utterance. The experiments are carried out with five noise classes namely babble, factory, vehicle, airport and hotel at various SNR thresholds and are analyzed using PESQ, WSS, Cepstral distance, frequency weighted segmented SNR and LLR. Significant improvement is noticed in the system performance [12].

Connected recognition and continuous Punjabi speech recognition is implemented. The DNN-HMM and GMM-HMM acoustic models are built using MFCC and GFCC features with mean and variance normalization of cepstral co-efficients. The LDA, MLLT, SAT, MLLR adaption methods are applied to achieve required level of dimensionality reduction [13].

Continuous Tamil speech recognition system is built for non-stationary noisy environmental conditions. A novel spectral subtraction speech enhancement method is explored by combining modified modulation magnitude estimation and Chi square distribution. The MFCC feature vectors are clustered using Fuzzy C-Means technique and are used to build FCM EM GMM, K-means EM GMM, FCM-HMM, K Means HMM acoustic models. The FCM EM-GMM outperforms all other models [14].

Long STM and Gated RNN architectures are explored for acoustic modelling of isolated words in Urdu language. The empirical analysis is carried out on various neural network architectures namely plain, bidirectional and deep directional neural networks. The model achieved 15% improvement in WER over baseline single layer LSTM. Further, boost in the performance is achieved by two layer LSTMs over Single Layer LSTMs but, the performance started decaying with further increment in LSTM layers. The LSTM architectures outperforms Gated Recurrent Unit Neural Network architectures [15].

DE optimized front features like MFCC, PLP, GFCC and MMI, MPE discriminative learning techniques are applied to continuous Hindi speech recognition. The performance is reported for various experimental approaches such as acoustic modelling context, modifying number of mixtures in the GMM, various optimized features with HMM, GMM-HMM, variations in the language models, using different

discriminative learning techniques i.e., GFCC optimized by DE with MMI, MPE, with clean and noise corrupted speech for various types of noises of changing SNR levels. The DE optimized GFCC coupled with MPE discriminative learning techniques with triphone language models outperformed all other model combinations in clean and noise affected conditions[16].

## III. METHODOLGY

### Preprocessing

The speech database developed in room environment is initially subjected to preprocessing operations like framing, windowing, pre-emphasis. A window duration of 20msec with an overlap of 10msec is considered for framing and windowing [17].

### Database

The speech sounds are recorded using Dell laptop at 16KHz sampling frequency and 16bits per samples ADC resolution. The isolated words database of 0-9 of English, Kannada language is recorded in room environment. The database consists of a 30 speakers uttering each digit 30 times. Thus creating a total of 300 isolated words. The 90% of the dataset is used for training and remaining 10% of the dataset is used for testing.

### MFCC Feature Extraction

The parameterization of speech is fundamental step for processing of speech. Speech signal is quasi-periodic signal[17]. The characteristics of speech remain stationary over a very short span of 10-20msec and the characteristics tend to change as the duration increases [17]. Thus, the short time Fourier analysis is very essential[17]. The Mel Frequency Cepstral coefficients are inspired by the human auditory filter banks[18]. The feature dimension is 39. The feature vectors are formed by appending delta and acceleration coefficients to first 13 MFCC coefficients. **3.3.1 Framing and pre-Emphases**

The long duration speech is bi-furcated as small duration segments based on the principle of time dependent processing of speech. To boost the energy of the speech wave the high frequency components are subjected to emphasis [19]. This is achieved using a difference equation [18] $s'_n = s_n - k.s_{n-1}$ (1)
Where $k = 0.97$ is the pre-emphasis co-efficient. Range of values that k takes are $0 \leq k < 1$ [18]

### Discontinuity Supression

The windowing step of MFCC process mainly aims for windowing speech signal using a Hamming window. The concept of windowing is employed to suppress the breaks in the continuity of speech signal specifically at the starting and ending points and to ensure the ending segments are suitable for connecting with the beginning portion of the segment [18]. This is accomplished by employing window function according to the equation [18].

$$s'_n = \{0.54 - 0.46\,.\,cos\left(\tfrac{2\pi(n-1)}{N-1}\right).s_n \qquad (2)$$

### Transformation of Waveform into spectrum

The speech waveforms each of N samples are mapped into spectrum using Fast Fourier Transform (FFT) [17]. The FFT is computationally efficient over Discrete Fourier Transform (DFT) [17].

$$S_n = \sum_{k=0}^{N-1} s_k\, e^{-j\frac{2\pi kn}{N}}, \quad n = 0,1,2,\dots,N-1 \qquad (3)$$

### Filter Bank Analysis

The experimental psychology portray that the human ear's resolution of frequency is not according to the linear scale across auditory spectrum [18]. So, for different values of frequencies, different subjective pitch values are obtained on Mel scale. The relationship between frequency and Mel frequency scale is linear below 1000Hz and is logarithmic above 1000Hz [19]. The triangular shaped filers are used here. The frequency to Mel transformation is according to the relation [18], [19]

$$Mel(f) = 2595\,.\,log_{10}\left(1 + \frac{f}{700}\right) \qquad (4)$$

### Discrete cosine transform (dct) and logarithm

The MFCC are extracted by logarithmically processing the filter bank amplitudes ($m_j$) using DCT [18]:

$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j\,.\,cos\left(\frac{\pi i}{N}(j - \frac{1}{2})\right) \qquad (5)$$

Here, N indicates the number of channels in the filter bank.

### Liftering

The fundamental usefulness of Cepstral domain coefficients is that they are related with each other [18]. But, higher order cepstral are relatively smaller, the cepstral coefficients are rescaled to ensure that all have similar magnitudes [18]. Liftering operation is used to fulfill this requirement according to the relation [18]. $c'_n = \left(1 + 0.5 * L.sin\left(\frac{\pi.n}{L}\right)\right).c_n$ (6)
Here, L denotes Cepstral sine lifter parameter.

### Plp Process

The parameterization of speech wav files into Perceptual Linear Prediction (PLP) coefficients is described in the upcoming paragraph.

### Spectrum Analysis

The first step here is to perform weighing operation on speech to segments by using Hamming Window [20]. The purpose of doing this operation is to ensure that the ending part of speech segment is smooth enough to connect with beginning part [17]. Every speech frame the speech is set to zero in the beginning and ending portions of the frame using Window by Eq. 2 and the DFT is computed by Eq.3

The square of real part and square of imaginary part are computed and summed to obtain the short-term power spectrum, according to the equation [20],

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2 \qquad (7)$$

**Critical-Band Spectral Resolution**

The mapping of frequency axis $\omega$ in $P(\omega)$ onto Bark frequency Ω is achieved by the equation [20]:

$$\Omega(\omega) = 6\ln\left\{\frac{\omega}{1200\pi} + \sqrt{\left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]}\right\} \quad (8)$$

$$\Omega(f) = 6\ln\left\{\frac{f}{600} + \sqrt{\left[\left(\frac{f}{600}\right)^2 + 1\right]}\right\} \quad (9)$$

$$\Omega(f) = 6sinh^{-1}\left(\frac{f}{600}\right) \quad (10)$$

Where $\omega$ is angular frequency, f is frequency in Hz. The warped power is convolved with the simulated critical-band masking curve's power spectrum $\Psi(\Omega)$ [20].

$$\Psi(\Omega) = \begin{Bmatrix} 0 & for & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & for & -1.3 \le \Omega \le -0.5 \\ 1 & for & -0.5 \le \Omega \le 0.5 \\ 10^{-1.0(\Omega-0.5)} & for & 0.5 \le \Omega \le 2.5 \\ 0 & for & \Omega > 2.5 \end{Bmatrix} \quad (11)$$

The power spectrum of critical band is obtained by convolving $\Psi(\Omega)$ with $P(\omega)$ using the relation [20]

$$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \quad (12)$$

The of a function of critical band masking curve and short time power spectrum reduces the Fourier domain resolution with reference to the original $P(\omega)$ [20]

**Equal-Loudness Pre-Emphasis**

The preemphasis is achieved by using equal-loudness curve, by employing the equation [20]:

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \quad (13)$$

The $E(\omega)$ is the approximation to the unequal sensitivity of the human auditory system at different values of frequency [20]-[21].

Where

$$E(\omega) = \frac{(\omega^2 + 56.8 * 10^6)\omega^4}{(\omega^2 + 6.3 * 10^6)^2 * (\omega^2 + 0.38 * 10^9)} \quad (14)$$

$$E(f) = \left[\frac{f^2}{f^2 + 1.6 * 10^5}\right] * \left[\frac{f^2 + 1.44 * 10^6}{f^2 + 9.6 * 10^6}\right] \quad (15)$$

**Intensity-Loudness Power Law**

All pole modeling is done after cubic-root amplitude compression. The relation below describes the power characteristics of the human ear [20].

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \quad (16)$$

**Autoregressive Modeling**

The all pole model is used to approximate $\Phi(\Omega)$. The poles of the model are estimated using the autocorrelation method. This method is called as Linear Prediction [20], [22]. The cepstral coefficients are obtained from LPC coefficients by converting these into frames of cepstra [20].

**Liftering**

The ensure that all cepstral coefficients have almost similar magnitude [17], the Liftering operation is performed using the equation (6)

**Acoustic Models**

The statistics of the speech features corresponding to each phoneme or word of the language are modeled using acoustic models. These models have analyze the feature vectors to extract the acoustic content. Acoustic models are built using probability distributions over acoustic space.

**Language Models**

The ASR is a mathematical problem and initially interpreted as statistical classification problem. Classes are described as sequence of words $W$. The features of the speech signal are used to derive the parameters denoted as $X$. The problem statement is then presented as the identifying the sequence of words which maximizes the probability $P(W/X)$ given as:

$$P\left(\frac{W}{X}\right) = \frac{P\left(\frac{x}{w}\right)P(W)}{P(X)} \quad (17)$$

Where $P(W)$ is the language model which is function of heavy constraints and the linguistic information about the words in the vocabulary. $P(X/W)$ is called as acoustic model.

**Recognition**

Recognition task is achieved using Hidden Markov Model (HMM)

**Hidden Markov Model**

Here we have used a 3-state and a 5-state Markov chain for determining the $P(W/X)$ The 3- state Markov chain is displayed in the figure 1 below

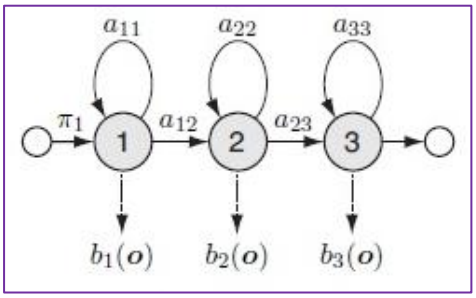


**Fig. 1 A 3-state left to right model**

During training stage, initial state probability ($\pi$), State transition probabilities (A), and output probabilities (B) have been computed using Baum-Welch Algorithm.

The HMM Model for every word is described by the relation

$$\lambda = (A, B, \pi) \quad (18)$$

The log-likely wood of every word is computed using Viterbi Decoding technique according to the equation

$$v* = [P\ (O|\lambda v)], \quad 1 \le v \le V \qquad (19)$$

$V$ is word length.

**Performance Analysis**

The performance of any ASR system is evaluated in terms of word error rate and word recognition accuracy [24] given by equations (20) and (21) respectively.

$$WER(\%) = \frac{(D+S+I)}{N} \times 100(\%) \qquad (20)$$

$$WRA(\%) = 100 - WER(\%) \qquad (21)$$

Where $N$ is the total number of words in the test speech corpus and $D, S\ and\ I$ are deletion, substitution and Insertion errors respectively. The above mentioned metrics are used for describing the performance of the proposed ASR system.

**Results Of Implemented Asr**

The ASR system is realized using HTK and Kaldi Toolkits. The performance is presented in terms of WER and WRA.

**Database**

The database created is classified into four sub classes, namely isolated digit corpus and continuous digit corpus. Each sub group has two subgroups, one for English Language and another for Kannada Language. For the ease of simplicity, we have labeled the speech corpus as indicated below.

- SC_GROUP1_IKD_SI
- SC_GROUP2_IED_SI
- SC_GROUP3_CKD_SI
- SC_GROUP4_CED_SI
- SC_GROUP5_IED_SI_UASPEECH DYSARTHRIC SPEECH DATABASE[25]

➢ The description about the Speech Corpus Group1 and Speech Corpus Group2 is as follows:

A thousand words speech corpus is created in room environment. The database consists of 50 repetitions of the English and Kannada digits through zero to nine spoken by 10 Male speakers aged between 25-32 years. Results of speaker independent digit recognition are presented in this section. Here, we have divided the database into train set and test set. The train set contains 450 words and test set contains 50 words. The system is built using Hidden Markov Model Toolkit (HTK version 3.4).

➢ The description about the Speech Corpus Group3 and Speech Corpus Group4 is as follows:

Speech Corpus Group3 and Group4 are speech data of continuously spoken digits through zero to nine and are collected by the 20 Male speakers and 10 female speakers for English and Kannada Language respectively. These two datasets are further bi-furcated into train and test sets.

The text pattern and corresponding Google transcription for data collection is as shown in Table 1. The ASR is implemented using Kaldi Toolkit. Performance of the system is evaluated using two metrics such as word recognition accuracy and word error rate.

➢ The description about the Speech Corpus Group5 is as follows:

Dysarthric speech data for UASPEECH is a corpus of Dysarthric speech created from 19 speakers having cerebral palsy. UASPEECH is developed by the University of Illinois. We have used three repetitions of isolated digits through zero to nine from the corpus for implementing the ASR [25]. The figure 2 is the general principle of ASR.

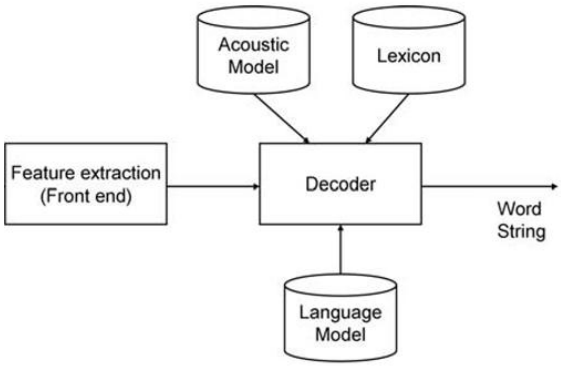


**Fig. 2 General principle of ASR**

**Table. 1 The text pattern and corresponding Google transcription for SC_GROUP3_CKD_SI data collection**

| Sl. No. | Text pattern and corresponding Google transcription |
|---|---|
| 1. | ಸೊನ್ನೆ ಒಂದು ಎರಡು ಮೂರು ನಾಲ್ಕು ಐದು ಆರು ಏಳು ಎಂಟು ಒಂಬತ್ತು<br>Sonne ondu eraḍu mūru nālku aidu āru ēḷu eṇṭu ombattu |
| 2. | ಒಂದು ಮೂರು ಐದು ಏಳು ಒಂಬತ್ತು ಸೊನ್ನೆ ಆರು ಎಂಟು ಎರಡು<br>Ondu mūru aidu ēḷu ombattu sonne āru eṇṭu eraḍu |
| 3. | ಸೊನ್ನೆ ಎರಡು ನಾಲ್ಕು ಆರು ಎಂಟು ಐದು ಒಂಬತ್ತು ಏಳು ಒಂದು<br>Sonne eraḍu nālku āru eṇṭu aidu ombattu ēḷu ondu |
| 4. | □□□□ □□□ □□□□□□□ □□□□ □□□□□□ □□□ □□□□□□ □□□<br>Ondu aidu ombattu eraḍu nālku āru sonne ēḷu |
| 5. | ಒಂದು ಮೂರು ಐದು ಆರು ಸೊನ್ನೆ ಎಂಟು ಒಂಬತ್ತು ಎರಡು<br>Ondu mūru aidu āru sonne eṇṭu ombattu eraḍu |
| 6. | ಎರಡು ಮೂರು ಒಂಬತ್ತು ನಾಲ್ಕು ಎರಡು ನಾಲ್ಕು ಐದು ಸೊನ್ನೆ ಏಳು ಆರು ಎಂಟು<br>Eraḍu mūru ombattu nālku eraḍu nālku aidu sonne ēḷu āru eṇṭu |

| | |
|---|---|
| **7.** | ಒಂದು ಮೂರು ಎರಡು ನಾಲ್ಕು ಐದು ಮೂರು ಸೊನ್ನೆ ಆರು ಒಂಬತ್ತು ಎಂಟು ಏಳು<br>Ondu mūru eraḍu nālku aidu mūru sonne āru ombattu eṇṭu ēḷu |
| **8.** | ಎರಡು ಮೂರು ಐದು ನಾಲ್ಕು ಎಂಟು ಏಳು ಒಂಬತ್ತು ಆರು ಒಂದು ಸೊನ್ನೆ<br>Eraḍu mūru aidu nālku eṇṭu ēḷu ombattu āru ondu sonne |
| **9.** | ಒಂಬತ್ತು ಎಂಟು ಏಳು ಆರು ಐದು ನಾಲ್ಕು ಮೂರು ಎರಡು ಒಂದು ಸೊನ್ನೆ ಏಳು ಒಂದು<br>Ombattu eṇṭu ēḷu āru aidu nālku mūru eraḍu ondu sonne ēḷu ondu |
| **10.** | ಮೂರು ನಾಲ್ಕು ಒಂದು ಸೊನ್ನೆ ಎಂಟು ಏಳು ಆರು ನಾಲ್ಕು ಮೂರು ಐದು ಎಂಟು ಒಂಬತ್ತು<br>Mūru nālku ondu sonne eṇṭu ēḷu āru nālku mūru aidu eṇṭu ombattu |

**Table. 2 Isolated digit recognition using SC_GROUP1_IKD_SI using 3-state HMM**

| GROUP1 | LPC | MFCC | PLP |
|---|---|---|---|
| **WRA** | 86.67 | 96.67 | 96.67 |
| **WER** | 13.33 | 3.33 | 3.33 |

**Table. 3 Isolated digit recognition using SC_GROUP2_IED_SI using 3-state HMM**

| GROUP2 | LPC | MFCC | PLP |
|---|---|---|---|
| **WRA** | 85.0 | 95.67 | 96.5 |
| **WER** | 15.0 | 4.33 | 3.5 |

**Table. 4 Isolated digit recognition using SC_GROUP2_IED_SI on 5-State HMM**

| GROUP2 | LPC | MFCC | PLP |
|---|---|---|---|
| **WRA** | 90 | 99.67 | 99.67 |
| **WER** | 10 | 0.33 | 0.33 |

**Table. 5 Continuous digit recognition of SC_GROUP3_CKD_SI Data using 3-State GMM-HMM (Mono), tri1, tri2 and tri3 using MFCC**

| GROUP3_MFCC | Mono | Tri1 | Tri2 | Tri3 |
|---|---|---|---|---|
| **WRA** | 94.0 | 95.75 | 96.0 | 98.25 |
| **WER** | 6 | 4.25 | 4.0 | 1.75 |

**Table. 6 Continuous digit recognition of SC_GROUP3_CKD_SI Data using 3-State GMM-HMM (Mono), tri1, tri2 and tri3 using PLP**

| GROUP3_PLP | Mono | Tri1 | Tri2 | Tri3 |
|---|---|---|---|---|
| **WRA** | 94.75 | 95.25 | 95.75 | 98.25 |
| **WER** | 5.25 | 4.75 | 4.25 | 1.75 |

The recognition results for isolated digit recognition of Kannada and Indian English numerals is presented in table 2 and table 3 respectively.

Table 4 denotes the recognition results for Indian English digits for a five state HMM. The continuous digit recognition performance of Kannada and English Language using MFCC and PLP with Monophone and tri-phone models is presented between Tables 5 to Table 8. The WER for UASPEECH is computed and shown in Table 9.

**Table. 7 Continuous digit recognition of SC_GROUP4_CED_SI Data using 3-State GMM-HMM (Mono), tri1, tri2 and tri3 using MFCC**

| GROUP4_MFCC | Mono | Tri1 | Tri2 | Tri3 |
|---|---|---|---|---|
| **WRA** | 92.75 | 95.0 | 95.25 | 97.0 |
| **WER** | 7.25 | 5.00 | 4.75 | 3.00 |

**Table. 8 Continuous digit recognition of SC_GROUP4_CED_SI Data using 3-State GMM-HMM (Mono), tri1, tri2 and tri3 using PLP**

| GROUP4_PLP | Mono | Tri1 | Tri2 | Tri3 |
|---|---|---|---|---|
| **WRA** | 92.75 | 95.75 | 96.25 | 96.75 |
| **WER** | 7.25 | 4.25 | 3.75 | 3.25 |

**Table. 9 The results for isolated digit recognition SC_GROUP5_IED_SI_UASPEECH DYSARTHRIC SPEECH DATABASE**

| GROUP5 | LPC | MFCC | PLP |
|---|---|---|---|
| **WRA** | 70 | 93.33 | 90.0 |
| **WER** | 30 | 6.67 | 10.0 |

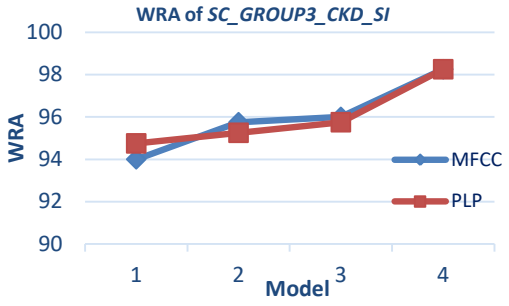


**Fig. 3 Word recognition accuracy of SC_GROUP3_CKD_SI for Monophone, tri1, tri2 and tri3 Models**

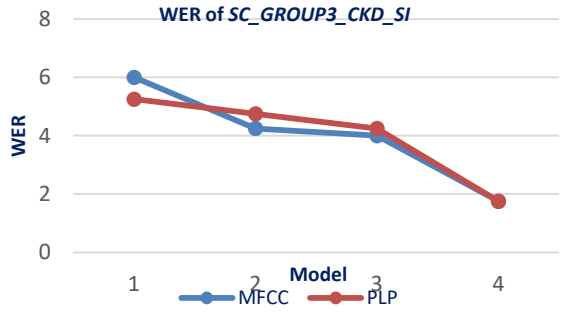


**Fig. 4 Word error rate of SC_GROUP3_CKD_SI for Monophone, tri1, tri2 and tri3 Models**

The figure 3 and figure 4 describe the WRA and WER for SC_GROUP3_CKD using MFCC and PLP. The figure 5 and figure 6 describe the WRA and WER for SC_GROUP4_CED using MFCC and PLP.

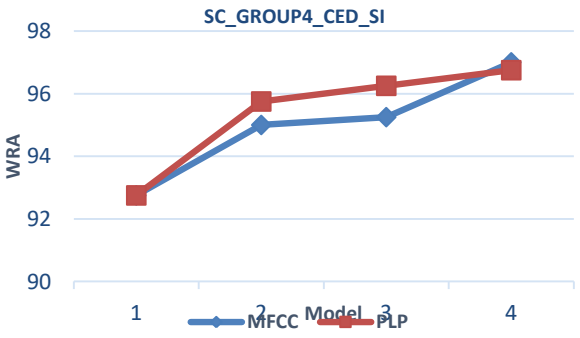


**Fig. 5 Word recognition accuracy of SC_GROUP4_CED_SI for Monophone, tri1, tri2 and tri3 Models**

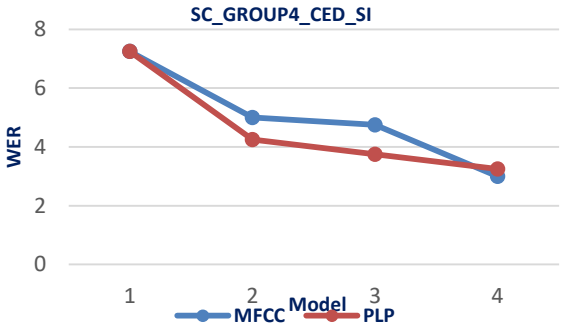


**Fig. 6 Word error rate of SC_GROUP4_CED_SI for Monophone, tri1, tri2 and tri3 Models**

**Table. 10 The text pattern and corresponding Google transcription for SC_GROUP4_CED_SI data collection**

| SL. NO. | Text pattern/corresponding Google transcription |
|---|---|
| 1. | Zero One Two Three Four Five Six Seven Eight Nine |
| 2. | One Three Five Seven Nine Zero Six Eight Two |
| 3. | Zero Two Four Six Eight Five Nine Seven One |
| 4. | One Five Nine Two Four Six Zero Seven |
| 5. | One Three Five Six Zero Eight Nine Two |
| 6. | Two Three Nine Four Two Four Five Zero Seven Six Eight |
| 7. | One Three Two Four Five Three Zero Six Nine Eight Seven |
| 8. | Two Three Five Four Eight Seven Nine Six One Zero |
| 9. | Nine Eight Seven Six Five Four Three Two One Zero Seven One |
| 10. | Three Four One Zero Eight Seven Six Four Three Five Eight Nine |

## IV. CONCLUSION

The performance of the proposed ASR systems are analysed in terms of WER and WRA.

- An Isolated digit recognition system is implemented using HTK over the three subgroup data sets. The system performance is reported for GMM-HMM framework over varied number of states of the HMM.
- The system is experimented with three groups of data sets. The acoustic information to feature representation is achieved using LPC, PLP and MFCC features.
- The system used the state of the art speech recognition classifier HMM with varied number of states.
- The use of MFCC and PLP lead to comparable performance, however both outperformed LPC features.
- The system built with 5-state HMM using MFCC outperformed the system realized using 3-state HMM with MFCC.

- A continuous digit recognition system is realised using Kaldi Toolkit over two subgroup data sets. PLP and MFCC features are utilized for realising tasks of ASR.
- The triphone3 system outperformed triphone2 system and triphone2 system performed slightly well than triphones1 and triphone1 outperformed monophone ASR system.

## REFERENCES


1. Long, Yanhua, Yijie Li, Hone Ye, and Hongwei Mao. "Domain adaptation of lattice-free MMI based TDNN models for speech recognition." International Journal of Speech Technology 20, no. 1 (2017): 171-178.
2. Rath, Shakti P. "Factored front-end CMLLR for joint speaker and environment normalization under DNN-HMM." International Journal of Speech Technology 20, no. 4 (2017): 859-867.
3. Kumar, Yogesh, and Navdeep Singh. "An automatic speech recognition system for spontaneous Punjabi speech corpus." International Journal of Speech Technology 20, no. 2 (2017): 297-303.
4. Maas, Andrew L., Peng Qi, Ziang Xie, Awni Y. Hannun, Christopher T. Lengerich, Daniel Jurafsky, and Andrew Y. Ng. "Building DNN acoustic models for large vocabulary speech recognition." Computer Speech & Language 41 (2017): 195-213.
5. Touazi, Azzedine, and Mohamed Debyeche. "An experimental framework for Arabic digits speech recognition in noisy environments." International Journal of Speech Technology20, no. 2 (2017): 205-224.
6. Londhe, Narendra D., and Ghanahshyam B. Kshirsagar. "Chhattisgarhi speech corpus for research and development in automatic speech recognition." International Journal of Speech Technology 21, no. 2 (2018): 193-210.
7. Guglani, Jyoti, and A. N. Mishra. "Continuous Punjabi speech recognition model based on Kaldi ASR toolkit." International Journal of Speech Technology 21, no. 2 (2018): 211-216.
8. Bharali, Sruti Sruba, and Sanjib Kr Kalita. "Speech recognition with reference to Assamese language using novel fusion technique." International Journal of Speech Technology(2018): 1-13.
9. Popli, Abhimanyu, and Arun Kumar. "Multilingual query-by-example spoken term detection in indian languages." International Journal of Speech Technology 22, no. 1 (2019): 131-141.
10. Manjunath, K. E., Dinesh Babu Jayagopi, K. Sreenivasa Rao, and V. Ramasubramanian. "Development and analysis of multilingual phone recognition systems using Indian languages." International Journal of Speech Technology 22, no. 1 (2019): 157-168.
11. Mittal, Puneet, and Navdeep Singh. "Development and analysis of Punjabi ASR system for mobile phones under different acoustic models." International Journal of Speech Technology 22, no. 1 (2019): 219-230.
12. Chiluveru, Samba Raju, and Manoj Tripathy. "Low SNR speech enhancement with DNN based phase estimation." International Journal of Speech Technology 22, no. 1 (2019): 283-292.
13. Kadyan, Virender, Archana Mantri, R. K. Aggarwal, and Amitoj Singh. "A comparative study of deep neural network based Punjabi-ASR system." International Journal of Speech Technology 22, no. 1 (2019): 111-119.
14. Kalamani, M., M. Krishnamoorthi, and R. S. Valarmathi. "Continuous Tamil Speech Recognition technique under non stationary noisy environments." International Journal of Speech Technology 22, no. 1 (2019): 47-58.
15. Zia, Tehseen, and Usman Zahid. "Long short-term memory recurrent neural network architectures for Urdu acoustic modeling." International Journal of Speech Technology 22, no. 1 (2019): 21-30.
16. Dua, Mohit, Rajesh Kumar Aggarwal, and Mantosh Biswas. "GFCC based discriminatively trained noise robust continuous ASR system for Hindi language." Journal of Ambient Intelligence and Humanized Computing 10, no. 6 (2019): 2301-2314.
17. Benba, Achraf, Abdelilah Jilbab, and Ahmed Hammouch. "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis." IEEE transactions on neural systems and rehabilitation engineering24, no. 10 (2016): 1100-1108.
18. Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore et al. "The HTK book." Cambridge university engineering department 3 (2002): 175
19. Benba, Achraf, Abdelilah Jilbab, and Ahmed Hammouch. "Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people." International Journal of Speech Technology 19, no. 3 (2016): 449-456
20. Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." the Journal of the Acoustical Society of America 87, no. 4 (1990): 1738-1752
21. Ho, Aileen K., Robert Iansek, Caterina Marigliani, John L. Bradshaw, and Sandra Gates. "Speech impairment in a large sample of patients with Parkinson's disease." Behavioural neurology 11, no. 3 (1999): 131-137
22. Logemann, Jeri A., Hilda B. Fisher, Benjamin Boshes, and E. Richard Blonsky. "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients." Journal of Speech and hearing Disorders 43, no. 1 (1978): 47-57
23. D. Povey, A. Ghoshal et. al, "The Kaldi Speech Recognition Toolkit", ASRU 2011
24. Upadhyaya, P., Farooq, O. & Abidi, M.R. Int J Speech Technol (2018) 21: 797. https://doi.org/10.1007/s10772-018-9545-2
25. Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame,Dysarthric Speech Database for Universal Access Research, Interspeech 2008, pp. 1741-4 (NSF 0534106; NIH DC008090A; Data)