Robust Perceptual Wavelet Packet Features for Recognition of Continuous Kannada Speech

Mahadevaswamy¹ · D. J. Ravi¹

Accepted: 9 July 2021 / Published online: 21 July 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
An ASR system is built for the Continuous Kannada Speech Recognition. The acoustic and language models are created with the help of the Kaldi toolkit. The speech database is created with the native male and female Kannada speakers. The 80% of collected speech data is used for training the acoustic models and 20% of speech database is used for the system testing. The Performance of the system is presented in terms of Word Error Rate (WER). Wavelet Packet Decomposition along with Mel filter bank is used to achieve feature extraction. The proposed feature extraction performs slightly better than the conventional features such as MFCC, PLP in terms of WRA and WER under uncontrolled conditions. For the speech corpus collected in Kannada Language, the proposed features shows an improvement in Word Recognition Accuracy (WRA) of 1.79% over baseline features.

Keywords Wavelet packet decomposition · Acoustic models · Hidden Markov Model · Deep Neural Networks

1 Introduction

The frequent pauses between the speech sounds of a speech signal portrays its unique characteristic that distinguishes it from all other signals. The speech database created in uncontrolled conditions of the environment must be processed to implement a robust automatic speech recognition system. Speech is an important and efficient tool of communication. The speech research drawn the attention of infinite researchers and has emerged as one of the important multidisciplinary research areas in the recent decades. Speaker Independent Speech recognition is the task of identifying the spoken word or sentence irrespective of the speaker. The speech recognition has been performed over the several languages. The UNESCO atlas of the world languages danger report-2009, describes that about 197 Indian languages are in critical situation of being extinct.
According to Indian census report, the percentage of people speaking local languages has drastically reduced [1]. A speech recognition system is implemented for Assamese language. The vocabulary size is 10 Assamese words. The task of speech recognition is achieved using Hidden Markov Model, I-vector technique and Vector quantization technique. A 39-dimensional features are derived using Mel Frequency Cepstral Coefficients, Delta-Coefficients, Acceleration Coefficients. The Novel Fusion technique outperforms the conventional techniques such as Hidden Markov Model, I-Vector Techniques and Vector Quantization Technique by achieving speech recognition accuracy of 100% [1]. The ASR system developed and evaluated using a moderate Bengali speech. Then 39-dimensional features are extracted and used to train triphone based HMM technique. The system was able to achieve an accuracy of 87.30% [2]. The speech recognition system is developed for Bangla accent. The Mel LPC features and their delta. The HMM modeling, lead to 98.11% recognition accuracy [3]. A Hindi isolated word recognition system is realized with LPC features and HMM Modeling and an accuracy of 97.14% was achieved corresponding to the word “teen” [4]. Another isolated word recognition system was realized with MFCC features and HTK Toolkit for Hindi language. An accuracy of 94.63% and a WER of 5.37% was achieved [5]. A connected word speech recognition system for Hindi language was proposed using MFCC features and HTK Toolkit. An accuracy of 87.01% was achieved [6].

An isolated digit recognition system was designed using MFCC features and HTK Toolkit for Malayalam isolated words to achieve an accuracy of 98.5% [7]. LPCC, MFCC, Delta-MFCC, Acceleration coefficients and vector quantization is utilized to build a speaker identification system to yield an accuracy of 96.59%. There is boost in the performance of the system by 3.5% accuracy during testing stage with a consideration to text dependent system [8]. An automatic language identification task is achieved among five Indian languages. The languages selected were are Hindi, Kannada, Telugu and Tamil. All the utterances are created from five native female speakers and five native male speakers. The cepstral features are derived from the speech signals and vector quantization technique based on the codebook concept is used to achieve the task of classification. The system achieved an recognition accuracy of 88.10% in recognizing spoken Kannada sentences [9]. A word recognition system was built for Punjabi language. The LPC feature vectors were extracted from speech signals. The vector quantization and Dynamic time warping techniques were used for implementing the speech recognition system. Experiments were carried out for different code book sizes from 8 to 256. The system was able to achieve a accuracy of 94% [10].

A speaker recognition system was developed for two speech databases. One speech database is created using microphone speech and other speech database is telephone speech. MFCC features are used with the Linear discriminant Analysis technique, Covariance Normalization, used to train the support vector machines classifier and cosine distance scoring [11]. The speech signal is a complex signal has information of vocal tract and the excitation source. To extract the excitation source information, the Linear Prediction Residual subjected to processing. The LP residual, Phase and Magnitude components are processed at three different levels, segmental level, sub-segmental level and suprasegmental level to derive the language specific excitation source information. The Gaussian Mixture Models are used to perform the classification task [12]. The literature reveals that the Kannada ASR system has not been experimented with Perceptual Wavelet Packet features so far. This approach is one of the first over Kannada language by augmenting the implementation of Perceptual Wavelet Packet features over the Kaldi toolkit. The organization of the article is as follows: In Sects. 1 and 2
provides introductory information towards automatic speech recognition and some of the important works presented in the literature. Section 3 describes about feature extraction methods.

2 Related Works

The automatic speech recognition (ASR) system is able to provide 100% accuracy under clean environment. But, its performance is degrades significantly when the spoken utterances gets contaminated by the presence of background noise or mismatch in acoustic features extracted from noisy or clean conditions [13–15] and mismatch in the labelled speech data used to train the classifier [16]. Hence, the performance of ASR system is constrained by two choices namely, correct labelling of speech data and selection of acoustic features. The well known acoustic features for speech recognition is Mel-frequency cepstral coefficients (MFCCs). MFCCs are extracted from the Mel filter banks[17]. MFCCs are obtained using short time Fourier transform (STFT). The mel cepstral coefficients are computed by allowing speech signal to pass through a bank triangular shaped filters having passbands slightly overlapping with adjacent passbands and to obtain a smooth spectrum [18, 19]. Spectrum is subjected variations as the impact of background noise increases [18, 20]. The popular MFCC technique consists of STFT. The STFT has a requirement that the signal to be processed must be stationary over short interval of time i.e.,semi-periodic signals[21]. Due to the trade-off between time–frequency resolution, it is not easy to detect phones that happen with a rapid burst in a slowly changing signal [18, 20, 22]. This problem of time–frequency resolution is alleviated by using wavelet transform(WT) [23–25]. The major benefit of using wavelet transform is that, unlike using single fixed sized analysis window in STFT, it uses windows with variable duration. The high frequency portion of the speech signal is processed by the short duration window, whereas the low frequency part of the speech signal is processed by the long duration window [24, 26, 27]. Thus by applying wavelet transform to a speech signal, it can be inspected for the presence or absence of sudden burst (stop phonemes) in a slowly changing signal [20, 22]. The conventional wavelet filter bank performed well for phoneme recognition tasks [20]. Because of the fixed resolution of frequency in time–frequency plane, the STFT was not able to find voiced stop due to their characteristic of rapid burst at higher frequencies [20, 22]. Multi-resolution potential of wavelets was enormously utilized by many research professionals for feature extraction and demonstrate their benefit for several applications such as, Biomedical application like ECG [28, 29], Speech enhancement [30, 31], EEG [32, 33] and Phoneme recognition [20, 22, 34].

3 Methodology

3.1 Preporocessing

The preprocessing functions like framing, windowing and pre-emphasis are applied to all the wave files in speech database. The frame duration and frame overlap are chosen as 20 ms and 10 ms respectively, for performing framing and windowing.
3.2 Proposed Features

The Multi-resolution property of the wavelet makes it appropriate tool for handling non-stationary and semi-stationary signals. This transform can detect unvoiced sounds in the speech signal and it provides best desnoising characteristics. In the recent years, several feature extraction approaches have been invented for speech recognition in uncontrolled environment. But, majority of these feature extraction schemes use Fourier transform to compute the spectrum. The speech signal consist of voiced (periodic) and unvoiced (aperiodic) portions throughout its existence. It’s a popular fact that the STFT or windowed Fourier transform has fixed and uniform frequency resolution with respect to the time frequency plane. Therefore it is difficult for the methods relay on STFT to recognize sudden bursts in the slowly time varying speech signals. To problem is alleviated by the application of wavelet transforms in the speech recognition research [35–37, 43–47]. The wavelet transform offers good frequency resolution[49–53, 55, 56].

3.2.1 Theoretical Background of Wavelet Transforms

Multi Resolution Analysis is an alternative way to STFT technique to analyze a signal. A mathematical scaling function is utilized to obtain a series of approximations to the signal. This principle has been considered by Wavelet Transforms (WT). A comparision of time-frequency resolution between STFT and WT is shown in Fig. 1.

3.2.2 Continuos Wavelet Transform (CWT)

CWT of a signal x(t) is given by

\[ CWT_{x}^{\psi}(\tau, s) = \frac{1}{\sqrt{S}} \int_{-\infty}^{\infty} x(t)\Psi^{*}(\frac{t-\tau}{s}) \, dt \]  (1)

From Eq. (1), the result of transformation is function of two variables, \( \tau \) and \( s \) that describe the translation and scaling factor respectively, and \( \Psi(t) \) is mother wavelet.

The term wavelet [38], is concatenation of two words ‘wave’ and ‘let’. Here wave is signal and let is short. The mother wavelet acts as a model or prototype to derive other

![Fig. 1 Comparision of STFT with WT](image)
window functions. The time information is captured by the variable $\tau$ and the parameter $s$ specifies dialation or compression operation on the wavelet.

### 3.2.3 Discrete Wavelet Transform (CWT)

The CWT is more complicated for signal analysis, because it involves significant computational resources. While DWT is less complicated in capture the signal information effectively [49]. The DWT of signal $x(t)$ is defined as:

$$DWT(j, k) = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} x(t) \psi\left(\frac{t - 2^j k}{2^j}\right) dt$$  \hspace{1cm} (2)

Mallat successfully demonstrated the method of wavelet decomposition by allowing a signal to pass through a series arrangement of low pass filter and high pass filter pairs. The multi resolution analysis of a signal is shown in Fig. 2a, b shown below. Here, $h_0(n), h_1(n)$ in the decomposition tree are low pass and high filter pairs respectively. Similarly $g_0(n), g_1(n)$ form the low pass and high pass filter pair in the reconstruction tree.

$h_0(n)\hspace{1cm}and\hspace{1cm}h_1(n)$ are a pair filters used for analysis, whereas $g_0(n)\hspace{1cm}and\hspace{1cm}g_1(n)$ form another pair of low, highpass filters respectively. These four filters have related as

$$h_1(n) = (-1)^n g_0(1 - n), \quad g_1(n) = (-1)^n h_0(1 - n)$$ \hspace{1cm} (3)

Also, the symbols $\downarrow^2$ and $\uparrow^2$ presented in Fig. 2a, b, denote the decimating and interpolating opertions carried out by a factor of 2 respectively. A pair of one level analysis and synthesis trees are shown in Fig. 3. In Fig. 3, $\{c_0(n)\} n \in \mathbb{Z}$ is the input applied the one level analysis and synthesis tree respectively [23].

$$c_1(k) = \sum_n h_0(n - 2k)c_0(n)$$ \hspace{1cm} (4)

$$d_1(k) = \sum_n h_1(n - 2k)c_0(n)$$ \hspace{1cm} (5)

where $c_1(k)$ and $d_1(k)$ are known as the approximation space and the detail space respectively. These are created by the one level wavelet analysis of $c_0(n)$. The corresponding synthesis tree is shown in Fig. 3 can be operated as

$$c_0(m) = \sum_k \left[ g_0(2k - m)c_1(k) + g_1(2k - m)d_1(k) \right]$$ \hspace{1cm} (6)

### 3.2.4 Wavelet Based Acoustic Feature Extraction

By repeating the iterative decomposition a desired binary wavelet packet tree is obtained. Various WP filterbank tree structures can be derived depending on application of interest. Wavelet features are extracted using Daubachies wavelet of order 4 (db4) [57]. Increasing the order of the mother wavelet may provide better results at expense of increased computational complexity.
Fig. 2  a The balanced 2-level analysis wavelet tree structure for $a_0$. b The balanced 2-level synthesis wavelet tree structure for $a_0$. 
3.2.4.1 Mel Filter Like WP Decomposition Farooq et al. [20], introduced 24-band Mel like Wavelet Packet Cepstral Features (WMFCC) The sound frequency $f_c$ is mapped to the mel frequency $f_{mel}$ according to the following equation

$$f_{mel} = 2595 \log_{10} \left( 1 + \frac{f_c}{700} \right)$$  \hspace{1cm} (7)$$

A frame size of 25 ms with a frame overlap of 15 ms was used to derive the WMFCC. Initially the speech frames are subjected to pre-emphasis followed by windowing operation using Hamming window. Initially a balanced three level wavelet packet tree structure is derived. Here, the frequency axis is subdivided into eight sub-bands each of 1 kHz The low frequency subband in the range 0–1 kHz is again subjected to three level balanced decomposition to get eight subbands each having a bandwidth of 125 Hz. Which is almost close to 100 Hz Mel-filter. The subband with frequency range is decomposed into two level balanced WP coefficients, giving four subbands each having a bandwidth of 250 Hz. The subbands in the range 1–1.25 kHz and 1.25–1.5 kHz are decomposed again, resulting in four subbands same bandwidth i.e., 250 Hz. The subband of 3-4 kHz frequency range is again processed by level decomposition, resulting in two subbands of 3–3.5 kHz and 3.5–4 kHz respectively. The frequency bands with ranges 4–5 kHz, 5–6 kHz, 6–7 kHz, and &–8 kHz are retained as it is. This results in 24-band Mel scale resembled WP filter. The bandwidth of the 24 frequency bands resulting after WP Decomposition does not exactly follow Mel scale [20] (see Table 1).

| Filters | Mel scale | Wavelet sub-band | Filters | Mel scale | Wavelet sub-band | Filters | Mel scale | Wavelet sub-band | Filters | Mel scale | Wavelet sub-band |
|---------|-----------|------------------|---------|-----------|------------------|---------|-----------|------------------|---------|-----------|------------------|
| 1       | 100       | 125              | 9       | 900       | 1125             | 17      | 2639      | 2750             |
| 2       | 200       | 250              | 10      | 1000      | 1250             | 18      | 3031      | 3000             |
| 3       | 300       | 375              | 11      | 1149      | 1375             | 19      | 3482      | 3500             |
| 4       | 400       | 500              | 12      | 1320      | 1500             | 20      | 4000      | 4000             |
| 5       | 500       | 625              | 13      | 1516      | 1750             | 21      | 4595      | 5000             |
| 6       | 600       | 750              | 14      | 1741      | 2000             | 22      | 5278      | 6000             |
| 7       | 700       | 875              | 15      | 2000      | 2250             | 23      | 6063      | 7000             |
| 8       | 800       | 1000             | 16      | 2297      | 2500             | 24      | 6954      | 8000             |

Fig. 3 The one level wavelet analysis and synthesis
The frequency axis is divided with the intention of matching it to frequency response of the Mel scale. The 24-band wavelet packet sub-bands resemble 24-band Mel filters [20] is shown in Fig. 4.

The energy in each subband is calculated by

\[ \langle S_i \rangle_k = \sum \frac{|\omega_p(x, k)|^2}{N_i} \]

where \( \omega_p(x, k) \) is wavelet packet coefficients of the signal \( x \), \( i \) is the subband frequency index \( 1 \leq i \leq M \), \( k \) indicates the temporal frame and \( N_i \) is the number of samples in the \( i^{th} \) subband. Similar to MFCC, the 24 energy coefficients are subjected to logarithmic compression. Finally, DCT is applied to all 24 coefficients and only first 13 normalized DCT coefficients are considered as WMFCC features. The pictorial representation of the feature extraction process is shown in Fig. 5.

3.2.4.2 Proposed PWP Tree Structure for Feature Extraction

In this work we have proposed a 24-band wavelet packet tree which is used to obtain the cepstral features. The feature extraction is carried out by proposing a 24-band Wavelet Packet (WP) tree structure after conducting repeated experiments. The WP tree structure shown in Fig. 6. is the proposed WP tree structure for obtaining the features.

The energy of the 24 band wavelet subbands are calculated. These coefficients are then logarithmically compressed and subjected to Discrete Cosine Transform. Discrete Cosine Transform (DCT) basically achieves energy compaction. The output of DCT gives 24 coefficients and only first 13 coefficients are used as cepstral coefficients. The Kaldi Toolkit is used to determine the delta and delta delta coefficients to form features of 39-dimension.
3.2.5 Acoustic Models

The acoustic models are used to map the observed feature matrix with the desired phoneme sequences of the hypothesized sentence. The creation of acoustic models is usually accomplished by using the Hidden Markov Models (HMM).

3.2.6 Language Models

The ASR systems utilize n-gram language models to facilitate the detection of exact word sequence through prediction of $n^{th}$ word, utilizing $(n - 1)$ previous words. Most popular n-gram language models are trigram ($n = 3$) and bigram ($n = 2$) language models.

3.2.7 Recognition

The speech recognition task is achieved using Gaussian Mixture Model–Hidden Markov Model (GMM–HMM), Triphones 1, Triphones 2 and Triphones 3 and Deep Neural Network (DNN).

3.2.8 Hidden Markov Model

To determine the probability $P\left(\frac{W}{X}\right)$ a 3-state Markov chain is used here. The 3- state Markov chain [54], is displayed in the Fig. 7. In the training phase, probability of system staying in a state ($\pi$), probability of transition between states ($A$), and output probabilities ($B$) are determined applying Baum–Welch Algorithm. The HMM Acoustic Model for each word sequence is defined by using the equation

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

(9)

The log-likelywood of every word sequence is estimated using Viterbi Decoding technique according to the equation

$$v = [P(O|\lambda v)], 1 \leq v \leq V$$

(10)
3.2.9 Performance Analysis

The recognition accuracy of any ASR system is determined using the popular metric word error rate and word recognition accuracy [24] given by Eqs. (11) and (12) respectively.

\[
WER(\%) = \frac{(D + S + I)}{N} \times 100(\%)
\]

\[
WRA(\%) = 100 - WER(\%)
\]

where \(N\) is the total number of words present in the test set and \(D, S, I\) are errors due to deletion, substitution and insertion respectively.

4 Database

The Kannada speech Database consisting of isolated digits from 0–9, 20 isolated words and 500 continuous speech sentences which includes combination of read speech and spontaneous speech and native language accents have been used as text script for creating the database. The database consists of 100 speakers. The continuous speech is of 15 h database. The database is recorded in the natural environment in the presence of room noise, vehicle noise happening nearby road at a distance of 5 m from the recording room. The tools used for recording the database is Matlab R2019b software on Dell Laptop. The speech data has been collected at a sampling rate of 16 kHz, 16-bit resolution and Mono recordings. The table indicates the 50 sample sentences from the speech database along with its google transcription. The lexicon is created using Indian Language Symbol Labels version 3 (ILSLv3 [39, 40]) prepared by ASR/TTS consortia, sponsored by Government of India, and hence have become the defacto standard; The continuous speech database created here
is according to the guidelines suggested by speech research expert Dr. Samudravijay K, from Indian Institute of Technology, Guwahati (IITG). The database is partitioned into training set and testing set. The 80% of the database is used for training the Acoustic Models and 20% of the database is used for testing the model.

The database consists of 3 sets for Kannada Language namely: isolated digits through (0–9), isolated words, Continuous Kannada Speech consisting of Spontaneous Spoken Kannada Sentences also. The database consists of 3 sets for English Language namely: isolated digits (TIMIT) through (0–9), isolated words (TIMIT), Librispeech of Continuous English Speech. The Kannada speech sentences along with their transcription is provided in the Table 2.

4.1 Kaldi Toolkit

Kaldi is an open source toolkit designed exclusively for building Acoustic Models (AMs) and Language Models (LMs) [41]. Kaldi is built using C++ programming language. The Kaldi toolkit can be run in windows as well as Linux based operating systems. But, the support for Linux based Kaldi tasks is very good compared to that of windows. The Table 3 presents the labels for kannada phones using syllable transliteration. There are four Dravidian languages. Kannada, Telugu, Malayalam and Tamil. Kannada is the most popular Dravidian language used in Karnataka state. This language consist of 14 swara (vowels), 32 vyanhana (consonants), 2 part vowel, yogavaahaka (part consonant). The labels used for building the lexicon for phonemes of the Kannada Language is shown in Table 2. Therefore, The Kannada language ASR system is developed by modeling the 46 phonemes. The labels are used from the Indian Speech Sound Label Set (ILSL12 [42]) is shown in Table 4. The lexicon for Kannada language is written by using ILSL12 label set shown in Table 5.

The dictionary is created by using ILSL12. Figure 8 represents the block diagram of the proposed features.

The general architecture of ASR system is shown in Fig. 9.

The Table 6 provides the details of parameters used for acoustic modelling. The acoustic models are generated at Monophones, Triphones1 and Triphones3 levels with number of jobs 3. The parameters used to develop Acoustic Model are as follows:

5 Results

The results of developed ASR system is presented in this section for Monophones, Triphones1, Triphones2, Triphones3 and DNN-HMM Phoneme Models. The Table 7 gives the WER details for the Kannada Isolated digit recognition task. The pictorial representation of Table 7 are presented in Fig. 10.

The Kannada isolated word recognition results and the corresponding graph are presented in Table 8 and Fig. 11 respectively.

The WER details for Continuous Kannada speech recognition for the speech data collected in uncontrolled conditions are presented in Table 9 and Fig. 12 respectively. In all the three sets of Kannada language a slight improvement in the performance can be observe with the proposed features over the MFCC and PLP features for DNN-HMM classifier.

The proposed features are also experimented with the isolated digits and isolated words extracted from the TIMIT database. The Table 10 and Fig. 13 describes the
| Kannada Text | Transliteration |
|--------------|-----------------|
| kannada text 1 | kannada transliteration 1 |
| kannada text 2 | kannada transliteration 2 |
| kannada text 3 | kannada transliteration 3 |

Table 2 The Kannada text sentences of the speech data collected and their corresponding transliteration.
Table 2 (continued)

| Feature                        | Performance improvement |
|--------------------------------|-------------------------|
| Robust Perceptual Wavelet Packet Features | Slight improvement       |

The Table 2 and Fig. 14 describe the performance of the proposed features over the MFCC and PLP features. A slight improvement in the performance can be observed in Table 11.

The proposed features are also experimented with standard Librispeech corpus of 08 h. A little improvement in the performance can be observed for the proposed features.
The Proposed ASR system is also tested with the unseen data consisting of 512 sentences of different combinations of words. A sample of 20 sentences are shown in the Table 13 and the results of the experiment are included in the Table 14.
Table 5  Dictionary for Kannada language is created by using ILSL12 label set

| TEXT TRANSCRIPTION     | LABEL SET USING ILSL12 |
|-----------------------|------------------------|
| koneyalli             | koneyallxi             |
| mattomme              | mattomme               |
| vaartegala            | vaartegala             |
| mukhyamshagalu        | mukhyanshagalx         |
| samsattina            | sannxattina            |
| ubhaya                | ubhaya                 |
| sadanagalalli         | sadanagalxi            |
| raastrapatigala       | raastrapatigala        |
| bhasanada             | bhasanxda              |
| meelina               | meelina                |
| vandana               | vandana                |
| nirnayada             | nirnxyada              |
| carce                 | carce                  |
| aarambha              | aarannxbha             |
| vaartegala            | vaartegalu             |
| vivara                | vivara                 |
| samsattina            | sanxsattina            |
| ubhaya                | ubhaya                 |
| sadanagalallindu      | sadanagalxinxdu        |
| raastrapatigala       | raastrapatigalx        |
| bhaasanakke           | bhasanxkkke            |
| vandane               | vandane                |
| sallisuvu             | sallxisuvu             |
| nirnayada             | nirnxyada              |
| meelina               | meelina                |
| carce                 | carce                  |
| aarambhavaagide       | aarannxbhavaagide      |

Fig. 8  Block diagram of proposed features
Fig. 9 ASR system architecture

Table 6 Parameters of THE ACOUSTIC MODEls

| Parameters specific to acoustic model | Triphone 1 | Triphone 2 | Triphone 3 |
|---------------------------------------|------------|------------|------------|
| Number of leaves                      | 2500       | 2500       | 2500       |
| Number of Gaussian                    | 20,000     | 20,000     | 20,000     |

Table 7 Kannada Isolated digit recognition

| SET1_DIGITS | Features- > | MFCC | PLP | Proposed |
|-------------|-------------|------|-----|----------|
| WER         | Mono        | 4.80 | 7.20| 20.00    |
|             | Tri 1       | 2.80 | 5.60| 10.40    |
|             | Tri 2       | 5.60 | 8.00| 8.80     |
|             | Tri 3       | 3.60 | 3.20| 2.80     |
|             | DNN-HMM     | 4.40 | 4.00| 2.80     |

Fig. 10 Comparision of WER for Kannada isoalted digit recognition over MFCC, PLP, proposed features
### Table 8  Kannada isolated word recognition

| SET2_WORDS | Features   | MFCC  | PLP  | Proposed |
|------------|------------|-------|------|----------|
| WER        | Mono       | 2.25  | 2.80 | 2.77     |
|            | Tri 1      | 1.30  | 2.46 | 3.08     |
|            | Tri 2      | 1.23  | 1.52 | 1.85     |
|            | Tri 3      | 0.92  | 1.30 | 1.23     |
|            | DNN-HMM    | 0.92  | 1.23 | 0.31     |

**Fig. 11** Comparison of WER for Kannada isolated word recognition over MFCC, PLP, proposed features

### Table 9  Continuous Kannada speech recognition using MFCC, PLP, Proposed features

| SET3_SENTENCES | Features   | MFCC  | PLP  | Proposed |
|----------------|------------|-------|------|----------|
| WER            | Mono       | 07.37 | 06.39| 16.20    |
|                | Tri 1      | 08.07 | 07.35| 08.96    |
|                | Tri 2      | 12.23 | 11.49| 14.09    |
|                | Tri 3      | 08.01 | 07.17| 07.01    |
|                | DNN-HMM    | 06.06 | 06.33| 04.27    |

**Fig. 12** Comparison of WER for Kannada continuous speech recognition over MFCC, PLP, proposed features
Table 10  English isolated digit recognition

| SET4_DIGITS     | Features | MFCC | PLP  | Proposed |
|-----------------|----------|------|------|----------|
| WER             | Mono     | 4.70 | 7.30 | 20.23    |
|                 | Tri 1    | 2.60 | 5.70 | 10.40    |
|                 | Tri 2    | 5.64 | 8.00 | 08.82    |
|                 | Tri 3    | 2.23 | 2.28 | 02.20    |
|                 | DNN-HMM  | 2.00 | 2.00 | 01.23    |

Fig. 13  Comparision of WERs for English isolated digit recognition over MFCC, PLP, proposed features

Table 11  English isolated word recognition

| SET5_WORDS     | Features | MFCC | PLP  | Proposed |
|----------------|----------|------|------|----------|
| WER            | Mono     | 1.50 | 4.25 | 4.75     |
|                | Tri 1    | 1.39 | 4.00 | 3.25     |
|                | Tri 2    | 1.31 | 3.50 | 1.50     |
|                | Tri 3    | 1.25 | 2.25 | 1.00     |
|                | DNN-HMM  | 1.01 | 1.00 | 0.75     |

Fig. 14  Comparision of WERs for English isolated word recognition over MFCC, PLP, proposed features
6 Conclusion

The ASR work carried out in this paper are as follows.

- We have experimented the conventional as well as proposed feature extraction technique over Monophone Models, Triphones1, Triphones2, triphones3 and DNN-HMM.
- The database consists of 3 sets for Kannada Language namely: isolated digits through (0–9), isolated words, Continuous Kannada Speech consisting of Spontaneous Spoken Kannada Sentences also.
- The database consists of 3 sets for English Language namely: isolated digits (TIMIT) through (0–9), isolated words (TIMIT), Librispeech of Continuous English Speech.
- In the experiments conducted over isolated digits and words taken from collected data of Kannada Language and from TIMIT data, the proposed features achieved significant improvement in the performance over the baseline features such as MFCC, PLP.
- For the experiments on collected Kannada Continuous Speech and Librispeech the proposed features are shown to perform better than the conventional features such as MFCC and PLP features.
- The Proposed ASR system is tested with the unseen data of 512 sentences and the performance on this test data set reveals that the proposed system performs better than the conventional features such as MFCC and PLP features.
| Sl. No. | Kannada Sentence                                                                 | Kannada Transcription                  | English Translation                  |
|--------|----------------------------------------------------------------------------------|----------------------------------------|--------------------------------------|
| 1.     | ನಮ್ಮ ಜಿವನದ ತಮ್ಮಬುಳ್ಳಾ ಕಾಸ್ತಾ                                                 | Nam’ma jīvanada tumb-a kaṣṭa           | The journey of our life is very difficult |
| 2.     | ಅರೋಗ್ಯ ಮುಖ್ಯಾದಾಯಕ ಕರ್ತೆ ಬಾರಿ                                              | Arōgya mukhya endu kalhe bari          | Tell the story that health is important |
| 3.     | ನಿನ್ನಾ ಮುಂದಿನ ಪ್ರಯಾಣ ಇನೆ ಇರಿಳಿ                                         | Ninna mundina prayaṇa ēnē irali        | whatever your next journey             |
| 4.     | ಅದರೆ ನಿಸೆಗಳು ಎಲ್ಲಕೈದು ಕಾಸ್ತಾ                                               | Adare ninage illē elakkinta kaṣṭa      | but it's hard for you here            |
| 5.     | ಅದಾರೆ ನಾಭೀ ನಿಸೆಗಳು ತಮ್ಮಿದ್ದಾಯಾನಿಯನಾದಾಯ ಮುಖ್ಯಾ                                         | Adare nambike ninage tumbā mukhya      | But faith is very important to you    |
| 6.     | ಮುಂದಿನ ಪ್ರಯಾಣ ಪ್ರಯಾಣನಿಯಾದ ಬಾರಿ ಬಾರಿ                                        | Mundina prayaṇa yāvudu endu bari       | Write about the next journey          |
| 7.     | ನಿನ್ನಾ ಮುಂದಿನ ಬಳಸಿ ಅನ್ಯಾ ಅಂಡೆ                                                | Ninna mundina kelasa ēnē irali        | Whatever your next job                |
| 8.     | ಅದಾರೆ ಬಾಡಾವರ ಜಿವನದ ಬಳಕೆ ಇಂದಿನ ಆರು                                            | Adare baḍavara jīvanada belaku ēnu     | But what is the light of the life of the poor |
| 9.     | ಅವರ ಜಿವಾಯಿ ಅಸತ್ಯ ಕತೆ                                                | Avara jīvana bari asatya kathe         | His life is simply untrue             |
| 10.    | ಅದಾರೆ ನಾಣೆ ಕತೆ ಇಂದಿನ ಮುಖ್ಯಾ                                              | Adare nanna kathē illē mukhya          | But my story is important here        |
| 11.    | ಕೌನ್ಯೈಲ್ಲ ಮಾತ್ತು ಕೋರುಮಾಟ್ಟ ವಿವರಾ                                               | Koneyalli mattonme vaartegala vivara    | At the end is the detail of the news once again |
| 12.    | ಮಾತ್ತು ಕೋರುಮಾಟ್ಟ ವಿವರಾ                                               | Mattomme samsattina sadanagalalli      | Once again in the houses of parliament |
| 13.    | ವಿವರಾ ತಾರದ ತಾರದ                                              | Vaartegala vivara aarambha             | The beginning of the news detail      |
| 14.    | ಭಾಸಾನದ ಮುಖ್ಯಾಮಾಧಾಪಳ ಪರುಸಿ ಪರುಸಿ                                                | Bhaasanada mukhyaamashagalu            | The main elements of talk             |
| 15.    | ಸಂಭಾವನೆ ಪ್ರತ್ಯೇಕ ಸಮುದ್ರವೆನಕೇರ ಪ್ರತ್ಯೇಕ ಸಮುದ್ರವೆನಕೇರ                              | Samsattina meolina nirayada carec      | The debate on the resolution on parliament |
| 16.    | ವಿವರಾ ತಾರದ ತಾರದ ತಾರದ                                              | Varthegala carec vivara aarambha       | Beginning of the discussion of the news |
| 17.    | ಸದನಗಾಲಿದು ಪ್ರತ್ಯೇಕ ಸಮುದ್ರವೆನಕೇರ ಪರುಸಿ                                             | Sadanagalallindu ubhaya samsattina     | Dual parliament in the house          |
| 18.    | ರಾಷ್ಟ್ರಪತಿಯ ಸಮುದ್ರವೆನಕೇರ ಸಮುದ್ರವೆನಕೇರ                                             | Raastrapatigala meolina vandane        | Salute to the president              |
| 19.    | ಭಾಸಾನಾಕ್ಕೆ ಸಳುಸುವಾ ವಿವರಾ                                              | Bhaasanakke sallisuva vivara          | Details to submit to the speech       |
| 20.    | ನಿರ್ಣಯದ ಕಾರೆ ಕಾರಾಂಬಹಾವೆಗಿದಿ ಕಾರಾಂಬಹಾವೆಗಿದಿ                                         | Nirayada carec aarambhavaagide         | The resolution debate has begun       |
Declaration

Conflict of interest  The authors declare that they have no conflict of interest.

References

1. Bharali, S. S., & Kalita, S. K. (2018). Speech recognition with reference to Assamese language using novel fusion technique. International Journal of Speech Technology, 21, 251. https://doi.org/10.1007/s10772-018-9501-1
2. Hassan, F., Khan, M. S. A., Kotwal, M. R. A., & Huda, M. N. Gender independent bangia automatic speech recognition. In International conference on informatics, electronics and vision (ICIEV-2012).
3. Muslima, U., & Islam, M. B. (2014). Experimental framework for melscaled LP based Bangla speech recognition. In 2013 IEEE 16th international conference on computer and information technology (ICCIT), Khulna (pp. 56–59).
4. Pruthi, T., Saksena, S., & Das, P. K. (2000). Swaranjali: Isolated word recognition for Hindi language using VQ and HMM. In International conference on multimedia processing and systems (ICMPS), Chennai.
5. Kumar, K., & Aggarwal, R. K. (2011). Hindi speech recognition system using HTK. International Journal of Computing and Business Research, 2(2), 2229–6166.
6. Kumar, K., Aggarwal, R. K., & Jain, A. (2012). A Hindi speech recognition system for connected words using HTK. International Journal of Computational Systems Engineering, 1(1), 25–32.
7. Kurian, C., & Balakrishnan, K. Speech recognition of Malayalam numbers. In IEEE World congress on nature & biologically inspired computing, 2009. NaBIC 2009, Coimbatore (pp. 1475–1479).
8. Bansal, P., Dev, A., & Jain, S. B. (2007). Automatic speaker identification using vector quantization. Asian Journal of Information Technology, 6(9), 938–942.
9. Balleda, J., Murthy, H. A., & Nagarajan, T. (2000). Language identification from short segments of speech. In Interspeech, Beijing.
10. Kumar, R., & Singh, M. (2011). Spoken isolated word recognition of Punjabi language using dynamic time warp technique. In C. Singh, G. S. Lehal, J. Sengupta, D. V. Sharma, & V. Goyal (Eds.), Information systems for Indian languages (pp. 301–301). Springer.
11. Senoussaoui, M., Kenny, P., Dehak, N., & Dumouchel, P. (2010). An I-vector extractor suitable for speaker recognition with both microphone and telephone speech. In Odyssey, Brno.
12. Nandi, D., Pati, D., & SreenivasaRao, K. (2017). Implicit processing of LP residual for language identification. Computer Speech and Language, 41, 68–87. https://doi.org/10.1016/j.csl.2016.06.002
13. Kim, C., & Stern, R. M. (2012). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4101–4104). IEEE. https://doi.org/10.1109/ICASSP.2012.6288820.
14. Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), 745–777. https://doi.org/10.1109/TASLP.2013.2304637
15. Mukherjee, H., Obaidullah, S. M., Santosh, K. C., Phadikar, S., & Roy, K. (2018). Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. International Journal of Speech Technology. https://doi.org/10.1007/s10772-018-9525-6
16. Bouguelia, M.-R., Nowaczyk, S., Santosh, K. C., & Verikas, A. (2017). Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *International Journal of Machine Learning and Cybernetics*, 9(8), 1307–1319. https://doi.org/10.1007/s13042-017-0645-0
17. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420
18. Farooq, O., Datta, S., & Shrotriya, M. C. (2010). Wavelet sub-band based temporal features for robust Hindi phoneme recognition. *International Journal of Wavelets, Multiresolution and Information Processing*, 08(06), 847–859. https://doi.org/10.1142/S0219691310003845
19. Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. In *Proceedings of the SPECOM* (pp. 191–194).
20. Farooq, O., & Datta, S. (2001). Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Processing Letters*, 8(7), 196–198. https://doi.org/10.1109/97.928676
21. Grigoryan, A. M. (2005). Fourier transform representation by frequency-time wavelets. *IEEE Transactions on Signal Processing*, 53(7), 2489–2497. https://doi.org/10.1109/TSP.2005.849180
22. Biswas, A., Sahu, P. K., Bhownik, A., & Chandra, M. (2014). Feature extraction technique using ERB like wavelet sub-band periodic and aperiodic decomposition for TIMIT phoneme recognition. *International Journal of Speech Technology*, 17(4), 389–399. https://doi.org/10.1007/s10772-014-9236-6
23. Biswas, A., Sahu, P. K., & Chandra, M. (2016). Admissible wavelet packet sub-band based harmonic energy features using ANOVA fusion techniques for Hindi phoneme recognition. *IET Signal Processing*, 10(8), 902–911. https://doi.org/10.1049/iet-spr.2015.0488
24. Steffen, P., Heller, P. N., Gopinath, R. A., & Burrus, C. S. (1993). Theory of regular M-band wavelet bases. *IEEE Transactions on Signal Processing*, 41(12), 3497–3511. https://doi.org/10.1109/78.258088
25. Vetterli, M., & Herley, C. (1992). Wavelets and filter banks: Theory and design. *IEEE Transactions on Signal Processing*, 40(9), 2207–2232. https://doi.org/10.1109/78.157221
26. Lin, T., Xu, S., Shi, Q., & Hao, P. (2006). An algebraic construction of orthonormal M-band wavelets with perfect reconstruction. *Applied Mathematics and Computation*, 172(2), 717–730. https://doi.org/10.1016/j.amc.2004.11.025
27. Pollock, S., & Ciasco, I. L. (2007). Non-dyadic wavelet analysis. In E. J. Kontoghiorghes & C. Gat (Eds.), *Optimisation, econometric and financial analysis* (pp. 167–203). Springer. https://doi.org/10.1007/3-540-36626-1_9
28. Chiu, C.-C., Chuang, C.-M., & Hsu, C.-Y. (2009). Discrete wavelet transform applied on personal identity verification with ECG signal. *International Journal of Wavelets, Multiresolution and Information Processing*, 07(03), 341–355. https://doi.org/10.1142/S0219691309002957
29. Rajoub, B., Alshamali, A., & Al-Fahoum, A. S. (2002). An efficient coding algorithm for the compression of ECG signals using the wavelet transform. *IEEE Transactions on Biomedical Engineering*, 49(4), 355–362. https://doi.org/10.1109/10.991163
30. Tabibian, S., Akbari, A., & Nasersharif, B. (2015). Speech enhancement using a wavelet thresholding method based on symmetric Kullback–Leibler divergence. *Signal Processing*, 106, 184–197. https://doi.org/10.1016/j.sigpro.2014.06.027
31. Zao, L., Coelho, R., & Flandrin, P. (2014). Speech enhancement with EMD and hurst-based mode selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Language Processing*, 22(5), 899–911. https://doi.org/10.1109/TASLP.2014.2312541
32. Adeli, H., Zhou, Z., & Dadmehr, N. (2003). Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods*, 123(1), 69–87. https://doi.org/10.1016/S0165-0270(02)00340-0
33. Ocak, H. (2009). Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. *Expert Systems with Applications*, 36(2), 2027–2036. https://doi.org/10.1016/j.eswa.2007.12.065
34. Biswas, A., Sahu, P. K., & Chandra, M. (2014). Admissible wavelet packet features based on human inner ear frequency response for Hindi consonant recognition. *Computers & Electrical Engineering*, 40(4), 1111–1122. https://doi.org/10.1016/j.compeleceng.2014.01.008
35. Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2), 171–185.
36. Gales, M. (2000). Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 8(4), 417–428.
37. Karpov, A., et al. (2014). Large vocabulary Russian speech recognition using syntactico-statistical language modeling. *Speech Communication*, 56, 213–228.
38. Daubechies, I. (1992). Ten lectures on wavelets. Society for industrial and applied mathematics.
39. http://www.iitg.ac.in/csweb/tts/assamese/Assamese%20Transliteration%20Tool%20(IT3%20to%20UTF-8)_11.html.
40. http://www.iitg.ac.in/samudravijaya/tutorials/ILSL_V3.2.pdf.
41. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., & Silovsky, J. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding, no. CONF. IEEE Signal Processing Society.
42. Yadava, T. G., & Jayanna, H. S. (2017). A spoken query system for the agricultural commodity prices and weather information access in Kannada language. International Journal of Speech Technology, 20, 635–644.
43. Mahadevaswamy, D. J. (2019). Performance of isolated and continuous digit recognition system using Kaldi Toolkit. International Journal of Recent Technology and Engineering, 8, 1–8.
44. Thimmaraja Yadava, G., & Jayanna, H. S. (2018). Creation and comparison of language and acoustic models using kaldi for noisy and enhanced speech data. International Journal of Intelligent Systems and Applications, 12, 22.
45. Praveen Kumar, P. S., Thimmaraja Yadava, G., & Jayanna, H. S. (2019). Continuous Kannada speech recognition system under degraded condition. Circuits, Systems, and Signal Processing, 39, 391–419.
46. Biswas, A., Sahu, P. K., Bhowmick, A., & Chandra, M. (2015). Hindi phoneme classification using Wiener filtered wavelet packet decomposed periodic and aperiodic acoustic feature. Computers & Electrical Engineering, 42, 12–22.
47. Mahadevaswamy, & Ravi, D. J. (2019). Performance of isolated and continuous digit recognition system using kaldi toolkit. International Journal of Recent Technology and Engineering, 8, 1–8.
48. Mahadevaswamy, & Ravi, D. J. (2016). Performance analysis of adaptive wavelet denosing by speech discrimination and thresholding. In 2016 international conference on electrical, electronics, communication, computer and optimization techniques (ICEECCOT), Mysuru (pp 173–178). https://doi.org/10.1109/ICEECCOT.2016.7955209.
49. Mahadevaswamy, & Ravi, D. J. (2016). Performance analysis of adaptive wavelet denosing by speech discrimination and thresholding. In 2016 international conference on electrical, electronics, communication, computer and optimization techniques (ICEECCOT), Mysuru (pp. 173–178). https://doi.org/10.1109/ICEECCOT.2016.7955209
50. Mahadevaswamy, & Ravi, D. J. (2017). Performance analysis of LP residual and correlation coefficients based speech separation front end. In 2017 international conference on current trends in computer, electrical, electronics and communication (CTCEEC), Mysore (pp. 328–332). https://doi.org/10.1109/CTCEEC.2017.8455039
51. Geethashree, A., & Ravi, D. J. (2018). Kannada emotional speech database: design, development and evaluation. In D. Guru, T. Vasudev, H. Chethan, & Y. Kumar (Eds.), Proceedings of international conference on cognition and recognition. Lecture notes in networks and systems. (Vol. 14). Springer.
52. Basavaiah, J., & Patil, C. M. (2020). Human activity detection and action recognition in videos using convolutional neural networks. Journal of Information and Communication Technology, 19(2), 157–183.
53. Basavaiah, J., & Anthony, A. A. (2020). Tomato leaf disease classification using multiple feature extraction techniques. Wireless Personal Communications. https://doi.org/10.1007/s11277-020-07590-x
54. Rajesh K Agrawal (2012). Improving Hindi Speech Recognition Using Filter Bank Optimization and Acoustic Model Refinement.
55. Mahadevaswamy (2021). Robust Continuous Kannada speech recognition using Kaldi toolkit.
56. Mahadevaswamy (2021). Voice Controlled IoT Based Grass Cutter Powered by Solar Energy.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Dr. Mahadevaswamy completed his B.E. degree from Visvesvaraya Technological University in the year 2010. M.Tech. in Digital Signal Processing from Jain University in the year 2012 and he is presently working as Assistant Professor in Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering (Autonomous), Mysuru. He completed his Ph.D. in the area of automatic speech recognition systems for Kannada Language. His area of interests include Artificial Intelligence, Machine Learning, Natural Language Processing, Digital Signal Processing and Speech Processing, and Communication.

Dr. D. J. Ravi received his B.E. degree in Electrical and Electronic Engineering in 1989 from the University of Mysore and M.Tech. in Industrial Electronic from Sri Jaya Chamaraja Engineering college in 1992, Mysuru and Ph.D. in Speech Processing from University of Mysore, India, in the year 2013. He is working as Professor in the department of electronics and communication engineering, Vidyavardhaka College of Engineering, Mysuru, India. His major research interest includes Speech synthesis and speech signal processing.