Pertanika J. Sci. & Technol. 26 (4): 1849 - 1866 (2018)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Performance Analysis of Isolated Speech Recognition System Using *Kannada* Speech Database

Ananthakrishna Thalengala[1]*, Kumara Shama[2] and Maithri Mangalore[3]

[1,2]Department of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal, India
[3]Department of Mechatronics, Manipal Institute of Technology, Manipal, India

ABSTRACT

In this article, performance analysis of speech recognition system for different acoustical models has been presented. In the present work, one of the well-known south Indian language named "*Kannada*" language is considered. Significantly large amount of work has been reported for Automatic Speech Recognition (ASR) in European languages whereas quite a small number of publications can be found in Indian languages. One of the reasons for this gap is that standard speech database in Indian languages is not available. In this study, *Kannada* speech corpus based on *Kannada* broadcast news data has been developed. The isolated speaker independent speech recognition system has been developed using Hidden Markov Tool Kit (HTK). The system front-end uses Mel frequency cepstral coefficients (MFCC) and its derivatives as acoustic features whereas acoustical models are developed by using Hidden Markov Models (HMM). Syllable and mono-phone based *Kannada* dictionaries have been developed in this study. Various mono-phone models considered in this work are word-level, syllable-level and phone-level models. Further, performance evaluation of mono-phone and tri-phone acoustical models for large sized dictionary also carried out. The best word recognition accuracies of 67.82% and 70.56% are reported for mono-phone and tri-phone based systems respectively. The recognition results for different HMM based acoustical models are obtained and hence the recognition performance has been analyzed.

ARTICLE INFO

Article history:
Received: 04 July 2016
Accepted: 17 April 2018
Published: 24 October 2018

E-mail addresses:
anantha.kt@manipal.edu (Ananthakrishna Thalengala)
shama.kumar@manipal.edu (Kumara Shama)
maithri.m@manipal.edu (Maithri Mangalore)
* Corresponding author

Keywords: Hidden Markov Tool Kit (HTK), *Kannada* language, Mel frequency cepstral coefficients (MFCC), Isolated Word Recognition (IWR) system, mono-phone model, phone dictionary, syllable dictionary, tri-phone model

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

One of the most natural and powerful means of human to human interactions is through speech communications. The human auditory perception system possesses natural ability to recognize speech with highest accuracy regardless of speech variations such as environmental noise, speaker's characteristics, emotions of the speaker, speaking rate and so on. And yet another speech variability is speaking accent which is based on the language or region. Even though there are significantly large number of research developments in ASR and state of the art speech recognition systems available, the system performances are not anywhere near to human auditory perception system. This is mainly because of the fact that speech is characterized by multi-modal variability such as emotion, speaker's characteristics, co-articulation effect, accent, and background noise. Among others, co-articulation effect and accent are more challenging variability to be addressed. It would be more appropriate to develop speech recognition system to specific application than looking for a generic applications. In this paper *Kannada* speech is considered with its applications in automatic name dialing (small vocabulary), railway information retrieval system (medium vocabulary) and broadcast news transcriptions (large vocabulary). A few of research contributions related to speech recognition for Indian languages are in Hindi (Kumar, Rajput, & Verma, 2004; Kumar, Aggarwal, & Jain, 2012; Saini, Kaur, & Dua, 2013), Tamil (Lakshmi & Murthy, 2006; Thangarajan, Natarajan, & Selvam, 2009; Radha, 2012), Telugu (Sunitha & Kalyani, 2012; Vijai Bhaskar, Rao, & Gopi, 2012), Marathi (Gawali, Gaikwad, Yannawar, & Mehrotra, 2011), Assamese (Bharali & Kalita, 2015), Punjabi (Dua, Aggarwal, & Kadyan, 2012) and *Kannada* (Hegde, Achary, & Shetty, 2012) (Punitha & Hemakumar, 2014; Hegde, Achary, & Shetty, 2015; Shridhara, Banahatti, Narthan, Karjigi, & Kumaraswamy, 2013).

LITERATURE REVIEW

In the recent few years, research in speech recognition for Indian languages is getting accelerated. Some of the research articles related to Hindi speech recognition are by Kumar, Rajput and Verma (2004), Kumar, Aggarwal and Jain (2012), and Saini, Kaur and Dua (2013). In these articles, authors have explored the speech recognition performance based on word-level, phoneme-level and tri-phone modelling methods. In most of the works authors have considered small or medium sized dictionary. The large vocabulary continuous speech recognition (LVCSR) system for Hindi reported a word recognition accuracy of 75% by using tri-gram language model (Kumar, Rajput, & Verma, 2004). A syllable based continuous speech recognition for Tamil language has been developed and shows that the accuracy of the proposed system is comparable with that of the baseline tri-phone system

(Thangarajan, Natarajan, & Selvam, 2009). A small vocabulary isolated word recognition system in Tamil has been reported a word recognition accuracy of 88% (Radha, 2012). Another work on syllable based speech recognition system for Telugu has reported speech recognition accuracy of 80% (Sunitha & Kalyani, 2012). Recently, Bharali and Kalita (2015) built an isolated word recognition system for Assamese language with vocabulary size of ten words and had reported maximum word recognition accuracy of 95%. In this work, performances of various speech parameters such as linear predictive coding (LPC) analysis, LPC cepstral coefficients (LPCEPSTRA), and Mel frequency cepstral coefficients (MFCC) have been explored. The 39-length MFCC (along with its derivatives) features have shown the best recognition accuracy. Panda & Nayak,(2016) had come out with new syllable segmentation technique for Indian languages, which would help to improve performance of syllable centric speech recognition systems. In this article authors have considered speech samples from three Indian languages, viz. Hindi, Odia, and Bengali.

The literature for speech recognition systems based on *Kannada* language speech corpus is very much limited. Hegde, Achary and Shetty (2012) had developed isolated speech recognition system for *Kannada* words using support vector machine (SVM) technique. The objective of the work was to explore SVM classifier for small vocabulary system. The SVM had been trained by using MFCC parameters of the speech and reported a word recognition accuracy of about 79% for a small vocabulary size of 10 *Kannada* words. Further, Hegde, Achary and Shetty, (2015) had extended their work towards the classification of 'alpha-syllabary' sounds in *Kannada* language. Alpha-syllabary units basically represent alphabets ('*Aksharas*') of *Kannada* language. Statistical analysis of acoustical features such as MFCC and LPC had been carried out and classification accuracies of individual vowels and consonants of *Kannada* using SVM and HMM classifiers had been explored. In another work by Shridhara, Banahatti, Narthan, Karjigi and Kumaraswamy (2013), a prosodically guided phonetic search engine for *Kannada* speech corpus had been implemented. Speech utterances in three different contexts namely, read mode, conversation mode, and extempore mode were considered in this work. The transcription of *Kannada* speech had been carried out in different layers and conventional speech recognition system using HTK had been built and phone recognition accuracies were analyzed. In our earlier study (Ananthakrishna, Maithri, & Shama, 2015; Thalengala & Shama, 2016), isolated speech recognition system based on *Kannada* speech corpus was presented. Comparative study of syllable based and phone based acoustical models were made for a medium sized vocabulary system. But in the present work, suitable sub-word level acoustical models for small, medium, and large vocabulary systems are proposed. Further, context-dependent tri-phone acoustical models are used to improve the recognition accuracy for the large vocabulary system.

The literature on speech recognition work related to the Indian languages are summarized and tabulated as shown in Table 1. The researchers have used different sub-word acoustical models to improve the system performance. A good word recognition accuracy (above 80%) has been reported in table 1 because of the small vocabulary size considered. Researchers have mainly explored on syllable and phone based acoustical models for different Indian languages such as *Hindi, Tamil, Telugu, Kannada*, *Assamese*, and *Bengali*. The majority of the Indian languages are syllable-timed and researchers have tried exploring this feature. The other challenge is to develop application specific speech database in a particular language as there are no standard database available. In the present work, three different set of speech corpus with vocabulary size of 10, 110 and 1498 words are developed. The small vocabulary consists of utterances of ten *Kannada* digits, the medium sized vocabulary is taken from *Kannada* short story and the large vocabulary is chosen from the *Kannada* broadcast news corpus. The entire speech database is developed by recording the speech utterances from 3 male and 3 female speakers. The objective of this study is to propose suitable sub-word level acoustical models for the different vocabulary based isolated word recognition systems. Also context dependent tri-phone based speech recognition system has been implemented and its performance against context independent mono-phone based systems has been analyzed. So, the novelty of this work mainly includes building of *Kannada* speech database, developing the phone and syllable based dictionaries, and performance evaluation of word recognition system for different vocabulary size.

Table 1
Summary of literature on IWR for Indian languages

Language/ Authors	Vocabulary	Number of Speakers	Acoustical Models Used	Accuracy (%)
Hindi (Aggarwal & Dave, 2011)	400 words	1	Gaussian mixture, Mono-phone HMM	80
Hindi (Saini, Kaur, & Dua, 2013)	113 words	9	Mono-phone HMM	90
Tamil (Radha, 2012)	50 words	10	Mono-phone HMM	88
Telugu (Sunitha, & Kalyani, 2012)	40 words	1	Syllable based acoustic modelling	80
Assamese (Bharali, & Kalita, 2015)	10 words	15	Word level HMM with different hidden states.	80
Kannada (Hegde, Achary, & Shetty, 2012)	10 words	5	SVM classifier	79
Kannada (Thalengala, & Shama, 2016)	1500 words	6	Syllable and phone based HMM	40 and 61

Kannada Language

India is basically multi-lingual country with more than 1500 spoken languages and about 150 languages have sizable speaking population. There are twenty-two official languages recognized by the Indian government and *Kannada* language is one of them. *Kannada* is one of the Dravidian languages of India (Krishnamurti, 2003) and is the state language of Karnataka state with the history of more than 2000 years. As per the recent survey there are about sixty million *Kannada* speaking peoples present inside and outside of Karnataka state. *Kannada* is a syllable-timed language having 52 basic alphabets (called "*Akshara Maala*") which are basically evolved from the "*Kadamba*" script. The alphabets of *Kannada* are grouped into three categories which are named as "*Swaragalu*" (vowels), "*Vyanjanagalu*" (consonants), and "*Yogavahakagalu*" (nasal like consonants) (Hegde, Achary, & Shetty, 2015; Krishnamurti, 2003). The organization of these "*Aksharas*" is well structured and arranged as per the place of articulations. The meaningful sequence of syllable-like alphabets gives rise to words in *Kannada* language. In *Kannada* language, the alpha-syllabary ('*Aksharas*') units are very stable and have unique pronunciations which are independent of their occurrence in a word or sentence. Researchers have tried to explore this characteristics in speech recognition systems (Hegde, Achary, & Shetty, 2012; Hegde, Achary, & Shetty, 2015).

Isolated Word Recognition (IWR) System

Any speech recognition system building has got two major stages viz. training stage and testing stage. The training phase consists of building acoustical models (HMM based sub-word models) from the input training speech data set and is as shown in Figure 1a. Figure 1b shows testing phase, where recognition of input testing data set is carried out by using the sub-word models generated in training stage.

The front-end of the speech recognition system comprises speech pre-processing and cepstral analysis operations which are common to both the training and testing phases. The initial pre-processing step involves pre-emphasis and speech framing operations. It is assumed that the speech is recorded in a controlled environment with minimum noise

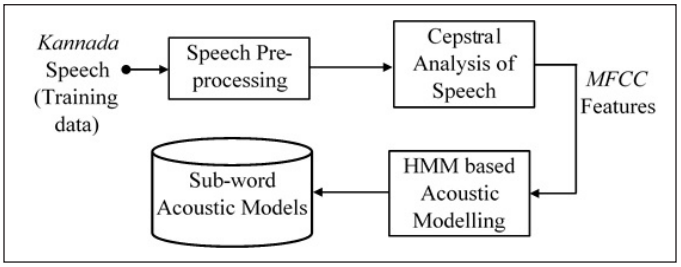

Figure 1a. Acoustic model building steps: The training phase

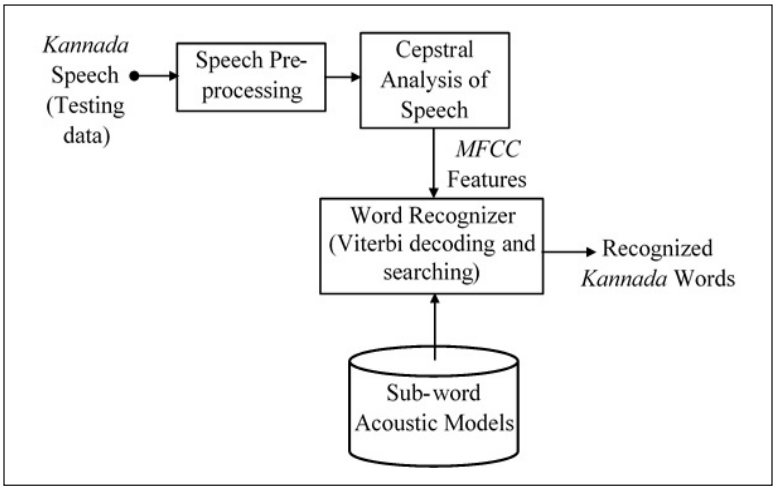

Figure 1b. Speech recognition steps: The testing phase

interference and uniform recording conditions. Speech signal has a characteristic that high frequency components have comparatively less energy than that of the low frequency components. This imbalance can be minimized by passing the speech signal through a pre-emphasis filter (Deller, Proakis, & Hansen, 1993). The spectrally flattened speech signal is segmented into overlapping time frames to obtain speech "frames" or "windows" and all the further processing is carried out on these speech frames. Speech is non-stationary in nature and therefore windowing is required to minimize the effect of non-stationary characteristics. The larger the speech frame length the better is the spectral resolution whereas the smaller the speech frame length the better is the time resolution. Typically window length of 1 or 2 pitch periods are considered in the speech analysis (Deller, Proakis, & Hansen, 1993); (Rabiner, Juang, & Yegnanarayana, 2012). In this work, hamming window of 20msec time duration and 10msec overlapping between the adjacent frames are considered.

The next stage in the front end of speech recognition system is speech parameterization. The performance of speech recognition system is very much dependent on the speech parameters considered in the study. Effectiveness of the speech parameters or features depends on how best it can represent acoustic models. Some of the speech features used in speech recognition systems are linear predictive coefficients (LPC), linear prediction cepstral coefficients (LPCC), perceptual linear prediction (PLP), and Mel frequency cepstral coefficients (MFCC). Among the others, MFCC features have shown better performance in speech recognition systems. These MFCC parameters have characteristics that it mimics the human ear perceptual system (Davis & Mermelstein, 1980). The response of human ear for different frequency bands is not linear but it can be well described by Mel-scale filter bank. In this study MFCC features together with its first and second order differences are considered.

Hidden Markov model (HMM) has been very commonly used in speech recognition task due to its ability to effectively model the time-series events. The back-end of the speech recognition system consists of building the HMM based acoustical sub-word units and applying pattern matching techniques to recognize the input utterances. The acoustical sub-word models considered in this study are words, syllables, and phone units of *Kannada* language. These acoustical sub-word units are developed using HMMs. The HMMs are basically the first order Markov process with non-observed states. So, HMM is a double stochastic process with one hidden state sequence which is calculated from the observation sequences. The HMM can be represented mathematically using its parameter set given by,

$$\lambda = (A, B, \pi) \qquad [1]$$

Where 'π' represents initial state distribution vector, '*A*' is state transition probability matrix and '*B*' is the state observation probability distribution vector (Rabiner, 1989; Nilsson, 2005). The most likely estimate of a word is computed by maximizing the a posteriori function according to,

$$\hat{W} = \arg\max_{W \in L} [P(O/W)P(W)] \qquad [2]$$

In the above equation [2], *(P(O|W))* is the probability of having certain observation sequence '*O*' of length '*L*' for a given word model '*W*'. *P(W)* represents the initial probability of a word model. Here observation vector is same as the feature vector and word model '*W*' consists of sequence of sub-word units. The summary of speech recognition process can be given by the following two steps.

Build individual HMM with same number of states for all the sub-word units. This is accomplished by using Baum-Welch algorithm and is known as HMM training.

1. Use the Viterbi decoding technique to find the best matching word. Decoding process uses the HMMs generated in the above step 1 to find the best state sequence for the given test sample. This is known as recognition or testing phase.
2. Finally, the performance of the system is evaluated based on the word recognition accuracy or using the word error rate (WER).

IWR System Implementation Using HTK

The isolated word recognition (IWR) system for *Kannada* words has been implemented using Hidden Markov Toolkit (Young et al., 2006) (HTK version 3.4.1) in the Linux platform. The system implementation involves mainly data preparation, data coding, acoustic modelling and evaluating the performance of the system. These steps are clearly presented in Figure 2.

In the data preparation stage, speech signal was acquired and then pre-processed. In addition to this, required vocabulary was defined and corresponding lexicon dictionary had been developed. In the speech analysis stage (data coding stage), speech signals were processed to obtain the sequence of feature vectors. The specifications such as window size, window type used, and choice of feature vector ere set during this stage. Next stage is the training phase, where speech features used to generate HMMs for each sub-word unit. Here Baum-Welch re-estimation algorithm was applied to obtain the acoustical sub-word models. This step is realized mainly by using the HTK commands "HCOMPV", "HREST", and "HEREST" as shown in Figure 2. Next, pattern matching of test samples with the stored acoustic models were carried out to obtain the recognized words. This is accomplished by using "HVITE" command in HTK. In the last stage, number of words classified correctly were computed and system recognition accuracy was analyzed.

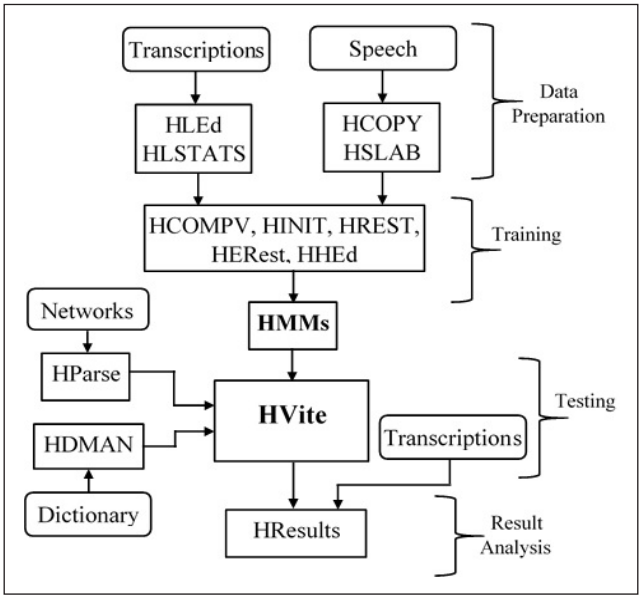

Figure 2. Steps in speech recognition system implementation using HTK

Speech Corpus

The details of speech databases developed for this study has been tabulated in table 2. There are three different speech databases considered in this study. A small vocabulary consisting of utterances of *Kannada* digits, a medium sized vocabulary developed by the word utterances from short *Kannada* story, and a large vocabulary obtained by recording the *Kannada* regional news corpus utterances. For instance, the number of training samples of 540 in Table 2 corresponds to the 10 words uttered by six different speakers nine times

each (10 words * 6 speakers* 9 times = 540 recordings). Similarly, the 180 test samples corresponds to the same 10 words uttered by the same six speakers three times. However the six speakers are different for different databases.

Table 2
Details of speech database developed

Speech Data Considered	Vocabulary Size (Number of Words)	Number of Speakers	Number of Training Samples	Number of Testing Samples
Kannada Digits (0 to 9)	10	6	540	180
Kannada Short Story	110	6	1980	660
Kannada Broadcast News	1498	6	5460	1820

These speech recordings have been carried out in a controlled (minimum noise interference) and uniform (good recording consistency) environment. A linear PCM recorder (Olympus LS-3) has been used in speech recordings and stored in wave format. Sampling frequency of 16 KHz and 16-bit PCM format has been followed while acquiring and digitizing the speech data.

Mono-phone Acoustical Models

The acoustical property and nature of sub-word units are highly influenced by the presence of adjacent sound units. This is mainly because of the co-articulation effect of the human speech production system. Two types of acoustical models can be defined based on the influence of adjacent sound units are context dependent tri-phone models and context independent mono-phone models. In the present study, three different sub-word mono-phone models and a tri-phone model based systems have been implemented. The tri-phone models are basically extended version of mono-phone models. The three mono-phone acoustical models considered in this study are word level model, syllable model, and phone model.

Developing any mono-phone model starts with building appropriate pronunciation dictionary. Both syllable and phone based dictionaries for the *Kannada* words have been developed. The phone set of *Kannada* language considered in this work are given in Table 3 (Shridhara, Banahatti, Narthan, Karjigi, & Kumaraswamy, 2013). There are 50 phones and are labelled with one or more English letters as shown in Table 3. The number of syllables (alpha-syllabary units) needed to represent chosen vocabulary size of 1498 words are found to be 564 syllables.

Syllables are obtained from the meaningful combination of phones and these syllables form alphabets of the *Kannada* language. Examples of how *Kannada* words are represented in pronunciation dictionaries (both phone and syllable based) are shown in Table 4. It

Table 3
Phones in Kannada language

Label	*Kannada* phone	Label	*Kannada* phone	Lable	*Kannada* phone
a	ಅ	g	ಗ್	p	ಪ್
A	ಆ	gh	ಘ್	ph	ಫ್
i	ಇ	ng	ಙ್	b	ಬ್
I	ಈ	c	ಚ್	bh	ಭ್
u	ಉ	ch	ಛ್	m	ಮ್
U	ಊ	j	ಜ್	y	ಯ್
ru	ಋ	jh	ಝ್	r	ರ್
rU	ೠ	ny	ಞ್	l	ಲ್
e	ಎ	T	ಟ್	v	ವ್
E	ಏ	Th	ಠ್	sh	ಶ್
Ai	ಐ	D	ಡ್	ss	ಷ್
o	ಒ	Dh	ಢ್	s	ಸ್
O	ಓ	N	ಣ್	h	ಹ್
Au	ಔ	t	ತ್	L	ಳ್
aom	ಅಂ	th	ಥ್	---	---
aH	ಅಃ	d	ದ್	---	---
k	ಕ್	dh	ಧ್	---	---
kh	ಖ್	n	ನ್	---	---

can be seen from the Table 4, that each *Kannada* word is represented as a sequence of *Kannada* syllables and sequence of *Kannada* phones. Clearly maximum number of phones required to represent the *Kannada* words is limited to 50, whereas number of syllables increases with vocabulary size up to a certain extent. The third acoustical model that has been considered in this study is word itself. This is used only for small vocabulary of 10 words (*Kannada* digit utterances). Word model has significantly poor performance and therefore not considered for larger databases.

Signal processing operations are performed on the speech waveform to obtain the cepstral parameters. It has been proven that Mel frequency cepstral coefficients (MFCC) are good parameters for representing acoustic units of speech (Davis & Mermelstein, 1980). Here speech waveforms are processed to obtain 12-length MFCC feature vectors. Along with the 12-MFCC parameters, signal energy component also included for the better representation. The human speech perception is greatly influenced by the spectral transitions present in the speech signal. This information can be quantified by considering the time difference (first order time derivatives) and acceleration coefficients (second order

time derivatives). So by combining all the parameters (12-MFCC, 1-log energy value, 13-delta coefficients, and 13-delta-delta coefficients) together the length of the feature vector becomes 39.

Table 4
Examples of phone and syllable sequences for Kannada words

Kannada word	English meaning	Sequences of syllables	Sequence of phones
ಸಹಕಾರ (sahakara)	Help	Sa ha kA ra	s a h a k A r a
ಮೂಲಕ (moolaka)	Through	mU la ka	m U l a k a
ಸಾವಿರ (savira)	Thousand	sA vi ra	s A v i r a
ಅವರು (avaru)	They	A va ru	a v a r u
ನೂತನ (noothana)	New	nU ta na	n U t a n a

Speech windows (Hamming window) of length 20 msec with 10 msec overlapping between the adjacent frames are considered in this study. The speech coding specifications used in HTK implementation (using "HCOPY" command) is tabulated as in Table 5. The speech parameterization is done for entire speech database by considering the both the training and testing speech samples. The feature vectors belonging to training set are used to develop the HMMs, whereas the one belonging to testing set are used to assess the models.

Table 5
Speech coding specifications

Parameters	Specifications
Sampling rate	16000 Hz
File format	16-bit PCM, mono, wave file
Window length	20 msec
Window used	Hamming
Window overlapping	50%
Pre-emphasis coefficient	0.97
Features derived	MFCC, ΔMFCC, and ΔΔMFCC
Feature vector length	39
Number of filter-banks	26
Number of MFCC coefficients	12

Context-independent mono-phone models have been built from the training feature set using the HMMs. This is implemented mainly by using "HCOMPV", "HREST", and "HEREST" commands in HTK tool as shown in figure 2 (Young et al., 2006). All the three acoustical models namely word-level, syllable-level and phone-level are considered in the design of mono-phone models. All the sub-word units are built by using 5-state HMM with first and last states are being non-emitting states. The HMM states are considered to be Gaussian distributed with first and last states representing the word boundaries (silence states). Initially the flat start prototype mono-phone models for each sub-word unit in the dictionary are defined. It is assumed that all the states in HMM are of Gaussian type with zero mean and unit variance. Features from the training data set are used to estimate and re-estimate the HMM based sub-word models. This is accomplished by the "HEREST" function, which adapts Baum-Welch re-estimation algorithm to obtain the trained HMMs. There are 50 phone models and about 500 syllable models have been developed in this study. The Baum-Welch re-estimation is a forward-backward algorithm which includes the pruning limit in its summation to reduce the amount of computations. Tight pruning thresholds can be kept for most of the training data sets, however, some training data set may show poor acoustic matching and as a result wider pruning threshold is needed. This issue is addressed by using an auto-increment pruning threshold. In the present study, pruning threshold of value of 250 has been considered, and if re-estimation fails, the threshold is incremented by a value 150. This is repeated until either the data is successfully processed or the pruning limit value of 1000 is exceeded.

After the training stage, recognition accuracy of the system evaluated using the testing data set. The Viterbi decoding technique ("HVITE" command in HTK) has been applied to calculate the best matching word from the dictionary for every testing sample. Based on the pattern matching results, recognition accuracy is obtained (using "HRESULTS" function).

Tri-phone Acoustical Models

The acoustical characteristics of a phone is highly influenced by the neighboring phones in any language. For instance, the utterance of the phone */a/* in the word "*acoustic*" sounds different for the same phone in the word "*automatic*". This is basically due to the co-articulation effects of the human speech production system. So, acoustical properties of phones are dependent on the occurrences of adjacent phones and hence context-dependent phones are to be defined. In tri-phone modeling, acoustical models for each phone is obtained by considering the effect of adjacent (immediate left and immediate right) phones. The tri-phone units are obtained from the phone sequences by concatenating the left and right phones. For a phone sequence "*th ih s*" (for the word "*this*"), the resultant tri-phone sequence would be "*th+ih th-ih+s ih-s*". The "+" and "-" symbols are used representing the right and left contexts of a phone respectively. Observe that the word boundaries gives

rise to bi-phone units. The conversion from mono-phone into tri-phone has got two steps. First, mono-phone transcriptions are cloned and converted into tri-phone transcriptions and tri-phone models are re-estimated. In the second step, the similar acoustic states of tri-phones are tied so that all state distributions are robustly estimated (Young et al., 2006). These implementation steps in HTK tool are summarized in Figure 3.

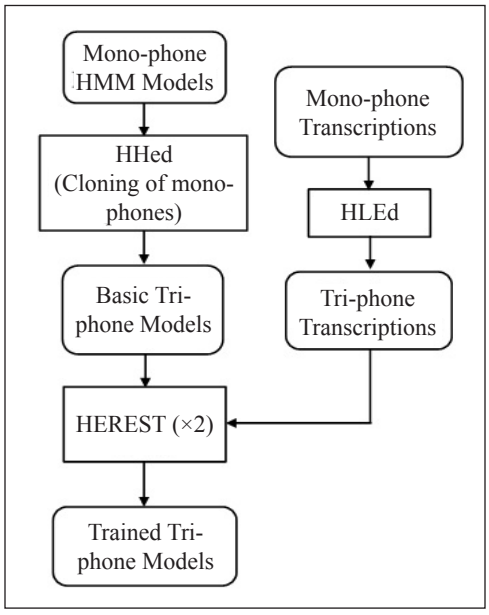

Figure 3. Mono-phones to Tri-phone conversion steps

In the previous step, the tri-phones generated such that the phone set share the same transition matrix. But the recognizer performance mainly depends on the accuracy of state output distributions that represent the input speech acoustics. So, the final stage in the tri-phone model building is to tie the states within tri-phones so that it is possible to make robust parameter estimation. This is achieved by 'HHED' command which clusters the states within tri-phones using decision tree method. The decision based technique requires questions to be set up regarding the left and right contexts of each tri-phones. Each question describes one set or class of contexts. For example a question named "left-stop-class" may define all possible "stops" to the left of any phone. The questions are defined such a way that they must be able to address the wide general class of sound units such as vowels, consonants, nasals and so on. Further, the questions can also include certain linguistic or phonetic classifications as appropriate. In the present tri-phone implementation the general class of phones are considered. But there is further scope to tune the tri-phones by exploring on the contextual occurrences of *Kannada* phone units. Once all the state-tying has been carried-out, the more refined tri-phone models are again re-estimated (using "HEREST" command) to obtain the final tri-phone models.

Finally, the word recognition was carried out by using Viterbi decoding algorithm (using "HVITE" command which is the same as given in Figure 2) on the test data set.

RESULTS AND DISCUSSION

The performance of any speech recognition system in general is evaluated by using word recognition accuracy. The word recognition accuracy is defined as follows.

$$\text{Recognition accuracy} = \frac{N-S}{N} \times 100\% \qquad [3]$$

Where '*N*' represents the number of words in the test data set and '*S*' is the number of words replaced (known as substitution error). Speech database considered here are obtained from the recordings of six different speakers and these databases have been pooled into three groups for experimentation. Each group has two speaker's data with one male and one female speaker. The well-known "*holdout*" procedure has been adapted in this analysis where one of the data group is used for testing and the remaining two for training. All the results presented here are the average of results for three data groups. The overall recognition result obtained in this study is summarized in Table 6. The performance analysis of the IWR system has been done based on the vocabulary size and on the acoustic model adapted. In this paper IWR system for small, medium, and large vocabulary are considered.

Table 6
Results of word recognition accuracy

Vocabulary Size	Acoustical Model Used	Number of HMMs	Accuracy (%)
10	word	10	97.22
10	Syllable	18	98.88
10	Phone	21	98.33
50	Syllable	93	96.33
50	phone	35	95.33
110	Syllable	138	95.15
110	Phone	37	94.24
921	Syllable	441	56.52
921	Phone	49	72.59
1498	Syllable	564	51.62
1498	Phone	49	67.82
921	Tri-phone	2114	82.57
1498	Tri-phone	3033	70.56

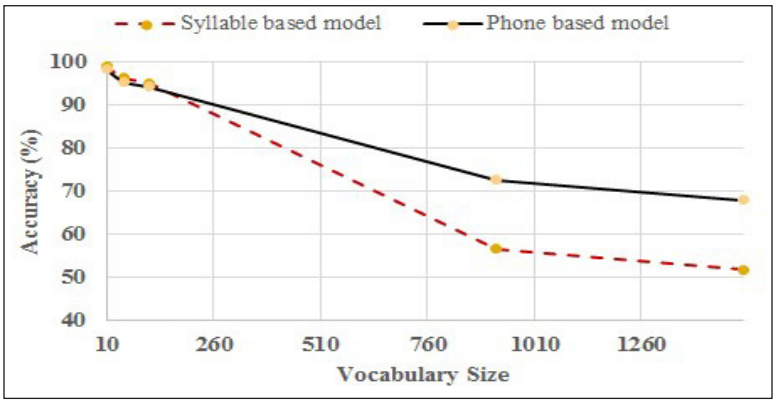

Figure 4. Performance of sub-word acoustical models

The syllable and phone based system performances for different vocabulary size has been plotted in Figure 4. It can be seen that the recognition accuracy of the both phone and syllable based systems decreases with the increase in vocabulary size. But the performance of phone based system is better than that of syllable based systems for larger vocabulary size. It can be observed that phone based system gives about 16% better results than the syllable based system for the vocabulary of 1498 words. The number of syllable models (number of HMMs) required increases as number of words increase, whereas number of phone models are limited to 50 for *Kannada* language. It is observed from the table 6 that forty nine phones are used and one phone left unused for the chosen vocabulary. Also when the acoustical models become larger in number, training that each model undergoes becomes lesser. Therefore we can see that the performance of syllable based system becomes poor with the increase in vocabulary size. Based on these findings, we propose the syllable models for small and medium sized vocabulary systems, whereas phone models for larger vocabulary systems. So the choice of acoustic model becomes application specific.

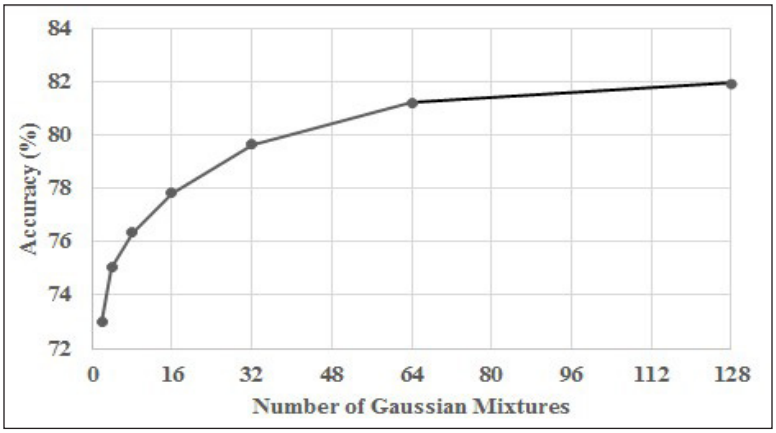

Figure 5. Performance of mono-phone models against number of Gaussian mixtures.

The individual states of HMM are assumed to be Gaussian distributed. For the better representation of sound unit, each HMM state can be made multivariate Gaussian density distributed function. But for the efficient modelling using Gaussian mixtures, a large training database is required. In this study, mono-phone models (phone based models) with Gaussian mixture models (GMM) have been implemented for vocabulary size of 1498 words. The results of recognition accuracy against the number of Gaussian mixtures is plotted in Figure 5. So the mono-phone system performance increases with the number of Gaussian mixtures up to a certain extent. But GMM implementation has a disadvantage that the training phase computational complexity is high and also require large training data set.

Now, all mono-phone models considered (word, syllable, and phone) in this work are context independent acoustic models. This means that each sub-word unit in a mono-phone system is built individually without considering its place of occurrence (context) in a word utterance. But the acoustical property of a sub-word unit in speech is greatly affected by the presence of adjacent sub-word units. In-order to address the phone-to-phone acoustical variations present in the utterance of a word, context dependent tri-phone models have been developed. The results of these tri-phone models are presented in the last two rows of Table 6. The results of context dependent tri-phone models show significant improvement over the mono-phone models. It can also be noted that the performance of GMM based mono-phone reaches near to the tri-phone performance with a mixture size greater than 64. Hence we propose to choose tri-phone models for large vocabulary *Kannada* word recognition systems.

CONCLUSION

Mono-phone and tri-phone based isolated word recognition systems for *Kannada* words have been realized in this study. The implementation of speech recognition system carried out successfully using HTK tool box. Context independent mono-phone systems with three different acoustical models namely word-level, syllable-level, and phone-level have been implemented for different vocabulary size. The implementation involves developing syllable and mono-phone dictionaries for *Kannada* words. Also context independent mono-phone models are extended towards context dependent tri-phone models. Recognition accuracies of these various implementations have been analyzed. The best word recognition accuracies of 67.82% and 70.56% are found respectively for mono-phone and tri-phone based systems on *Kannada* broadcast news database having vocabulary size of 1498 words.

It can be therefore proposed that syllable based acoustical models are suitable for small and medium size vocabulary systems. For applications requiring larger vocabulary, phone based models or context dependent tri-phone models can be chosen. System performance may be improved further by increasing the training data set. Also, there is a further scope for tuning the tri-phone models by incorporation language specific classifications in the questions tree.

REFERENCES

Aggarwal, R. K., & Dave, M. (2011). Using Gaussian mixtures for Hindi speech recognition system. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 4*(4), 157-170.

Ananthakrishna, T., Maithri, M., & Shama, K. (2015, December). *Kannada* word recognition system using HTK. In *2015 Annual India Conference (INDICON)* (pp. 1-5). New Delhi, India.

Bharali, S. S., & Kalita, S. K. (2015). A comparative study of different features for isolated spoken word recognition using HMM with reference to Assamese language. *International Journal of Speech Technology, 18*(4), 673-684.

Bhaskar P. V., Rao R. M. S. & Gopi A. (2012). HTK Based Telugu Speech Recognition.*-International Journal of Advanced Research in Computer Science and Software Engineering, 2*(12), 307-314.

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.

Deller J. R., Proakis J. G. & Hansen J. H. L., (1993). *Discrete Time Processing of Speech Signals*. New York: Macmillan Publishing Company.

Dua, M., Aggarwal, R. K., Kadyan, V., & Dua, S. (2012). Punjabi automatic speech recognition using HTK. *IJCSI International Journal of Computer Science Issues, 9*(4), 1694-0814.

Gawali, B. W., Gaikwad, S., Yannawar, P., & Mehrotra, S. C. (2011). Marathi isolated word recognition system using mfcc and dtw features. *ACEEE International Journal on Information Technology, 1*(01), 21-24.

Hegde, S., Achary, K. K., & Shetty, S. (2012). Isolated word recognition for *Kannada* language using support vector machine. In *Wireless Networks and Computational Intelligence* (pp. 262-269). Berlin, Heidelberg: Springer.

Hegde, S., Achary, K. K., & Shetty, S. (2015). Statistical analysis of features and classification of alphasyllabary sounds in *Kannada* language. *International Journal of Speech Technology, 18*(1), 65-75.

Krishnamurti, B. (2003). *The Dravidian Languages*. Cambridge: Cambridge University Press.

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.

Kumar, M., Rajput, N., & Verma, A. (2004). A large-vocabulary continuous speech recognition system for Hindi. *IBM Journal of Research and Development*, *48*(5.6), 703-715.

Lakshmi, A., & Murthy, H. A. (2006). A syllable based continuous speech recognizer for Tamil. In *Ninth International Conference on Spoken Language Processing* (pp. 1878-1881). Pittsburgh, Pennsylvania.

Nilsson, M. (2005). *First Order Hidden Markov Model: Theory and implementation issues*. Technical Report 2005:02. Blekinge Institute of Technology.

Panda, S. P., & Nayak, A. K. (2016). Automatic speech segmentation in syllable centric speech recognition system. *International Journal of Speech Technology, 19*(1), 9-18.

Punitha, P., & Hemakumar, G. (2014, March). Speaker dependent continuous *Kannada* speech recognition using HMM. In *2014 International Conference on Intelligent Computing Applications (ICICA),* (pp. 402-405). Coimbatore, India.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Rabiner, L. R., Juang B. H., & Yegnanarayana B. (2012). *Fundamentals of speech recognition*. Englewood Cliffs: PTR Prentice Hall.

Radha, V. (2012). Speaker independent isolated speech recognition system for Tamil language using HMM. *Procedia Engineering, 30*, 1097-1102.

Saini, P., Kaur, P., & Dua, M. (2013). Hindi automatic speech recognition using HTK. *International Journal of Engineering Trends and Technology (IJETT), 4*(6), 2223-29.

Shridhara, M. V., Banahatti, B. K., Narthan, L., Karjigi, V., & Kumaraswamy, R. (2013, November). Development of *Kannada* speech corpus for prosodically guided phonetic search engine. In *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference* (pp. 1-6). Gurgaon, India.

Sunitha, K. V. N., & Kalyani, N. (2012). Isolated Word Recognition using Morph Knowledge for Telugu Language. *International Journal of Computer Applications, 38*(12), 47-54.

Thalengala, A., & Shama, K. (2016). Study of sub-word acoustical models for *Kannada* isolated word recognition system. *International Journal of Speech Technology, 19*(4), 817-826.

Thangarajan, R., Natarajan, A. M., & Selvam, M. (2009). Syllable modeling in continuous speech recognition for Tamil language. *International Journal of Speech Technology, 12*(1), 47-57.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X & Valtchev, V. (2006). *The HTK book*. Cambridge, UK: Cambridge University Press.