

Study of sub-word acoustical models for *Kannada* isolated word recognition system

Ananthakrishna Thalengala¹  · Kumara Shama¹

Received: 10 June 2016 / Accepted: 22 September 2016 / Published online: 1 October 2016
© Springer Science+Business Media New York 2016

Abstract The speech recognition system basically extracts the textual information present in the speech. In the present work, speaker independent isolated word recognition system for one of the south Indian language—*Kannada* has been developed. For European languages such as English, large amount of research has been carried out in the context of speech recognition. But, speech recognition in Indian languages such as *Kannada* reported significantly less amount of work and there are no standard speech corpus readily available. In the present study, speech database has been developed by recording the speech utterances of regional *Kannada* news corpus of different speakers. The speech recognition system has been implemented using the Hidden Markov Tool Kit. Two separate pronunciation dictionaries namely phone based and syllable based dictionaries are built in-order to design and evaluate the performances of phone-level and syllable-level sub-word acoustical models. Experiments have been carried out and results are analyzed by varying the number of Gaussian mixtures in each state of monophone Hidden Markov Model (HMM). Also, context dependent triphone HMM models have been built for the same *Kannada* speech corpus and the recognition accuracies are comparatively analyzed. Mel frequency cepstral coefficients along with their first and second derivative coefficients are used as feature vectors and are computed in acoustic front-end

processing. The overall word recognition accuracy of 60.2 and 74.35 % respectively for monophone and triphone models have been obtained. The study shows a good improvement in the accuracy of isolated-word *Kannada* speech recognition system using triphone HMM models compared to that of monophone HMM models.

Keywords *Kannada* language · Hidden Markov Tool Kit (HTK) · Hidden Markov Model (HMM) · Mel frequency cepstral coefficients (MFCC) · Phone-level HMM · Syllable-level HMM · Triphone model

1 Introduction

Speech recognition basically deals with transcription of speech into text in a given language. Even today, developing an efficient speech recognition system is challenging task due to different variabilities present in the speech signal. Some of the prominent factors that affect the recognition accuracy are speaker to speaker variations, rate of speech utterance, mood of the speaker, recording environment, vocabulary size, nature of speech utterance (continuous or isolated words) and so on. However, automatic speech recognition (ASR) systems for European languages such as English, French and German are well developed. In Indian sub-continent there are more than 200 written languages with 22 government recognized official languages. Speech sounds in each language differ by the phonetic structure and variable vocabulary size. Most of the Indian languages are syllable-timed (Panda and Nayak 2015) and there exist one to one mapping between written script and its pronunciation. So, sub-word level acoustical models are to be explored in the back-end design of ASR system for Indian languages to obtain better results. The front-end of

✉ Ananthakrishna Thalengala
anantha.kt@manipal.edu

Kumara Shama
shama.kumar@manipal.edu

¹ Department of Electronics and Communication Engineering,
Manipal Institute of Technology (MIT), Manipal 576104,
India

ASR system consists of signal processing operations to extract relevant features and MFCCs along with their derivatives which are proved to be robust features for representing the speech sub-word units (OShaughnessy 2008).

1.1 Literature overview

In the past few years, speech recognition for Indian languages is gaining lot of importance. Researchers in the speech processing domain have done significant amount of contribution in different Indian languages such as Hindi (Aggarwal and Dave 2011, 2012; Kumar and Aggarwal 2012; Saini et al. 2013; Neti et al. 2002), Tamil (Thangarajan et al. 2009; Lakshmi and Murthy 2006; Radha et al. 2012), Telugu (Mannepalli et al. 2016; Sunitha et al. 2012; Bhaskar et al. 2012), Assamese (Bharali and Kalita 2015), Bengali (Hassan et al. 2011) and Kannada (Hegde et al. 2015, 2012; Hemakumar and Punitha 2014a). Neti et al. (2002) of IBM India research lab, have developed a large-vocabulary continuous speech recognition system for Hindi. Also, Kumar and Aggarwal (2012), Aggarwal and Dave (2011), Aggarwal and Dave (2012), Saini et al. (2013), and Mishra et al. (2011) have contributed to speech recognition in Hindi language. In these research publications, authors have analyzed the recognition results based on word-level, phoneme-level and triphone modeling methods for a medium vocabulary systems.

In Tamil language, Lakshmi and Murthy (2006) presented a novel technique for building a syllable based continuous speech recognizer and Thangarajan et al. (2009) have developed a small vocabulary word-based and medium vocabulary triphone based continuous speech recognition system. Recently, Radha et al. (2012) reported work on small vocabulary, isolated speech recognition system for Tamil language. The system performance in terms of word recognition accuracy of about 88 % has been reported. The researchers in Telugu language (Sunitha et al. 2012) have built syllable based recognition system and have achieved good word recognition accuracy of about 80 %. The system was trained with a vocabulary size of 400 words and tested using another 100 words. Also, in 2012, Bhaskar et al. (2012) have built speech recognizer for Telugu language using the HTK toolkit. Recently, Mannepalli et al. (2016) have developed Gaussian mixture model (GMM) for Telugu dialect recognition based on MFCC features. Bharali and Kalita (2015) have done a comparative study of recognition accuracy of Assamese spoken digits for different feature sets and for various number of states in HMM. An automatic speech recognition system based on context sensitive triphone acoustic models has been implemented for Bengali language (Hassan et al. 2011). Here, the proposed method comprises of

extracting phoneme probabilities using multilayer neural network (MLN), designing context sensitive triphone models and generating word strings based on triphone HMMs.

The number of speech recognition works reported for Kannada language is significantly less. Hegde et al. (2012) developed an isolated word recognizer for identification of spoken words for the database created by recording the words in Kannada language. In this work, support vector machine (SVM) algorithm has been used for designing the classifier model. Recently, the same authors have reported a work on classification of alphasyllabary sound units of Kannada language (Hegde et al. 2015). In this work, authors have comparatively analyzed the performances of SVM and HMM classifiers for Mel frequency cepstral coefficients (MFCC) and linear predictive coefficients (LPC) feature sets. Using short-time energy and signal magnitude, Hemakumar and Punitha (2014a) have done segmentation of Kannada speech into syllable-like sub-word units. Also, the same authors have reported a work on speaker dependent Kannada speech recognition system based on discrete HMM (Hemakumar and Punitha 2014b). In fact, in our earlier work (Muralikrishna et al. 2013; Ananthakrishna et al. 2015), we have made a similar attempt on Kannada spoken digit recognition system by considering MFCCs along with their first and second order derivatives as the feature vectors.

The survey on the recent literature on speech recognition in Indian languages shows that researchers try to explore the different sub-word modeling techniques. But majority of the works are limited to medium size (maximum of 100 words) vocabulary system. As most of Indian languages are syllable timed, the syllable-based acoustic models have outperformed against the phone-based counterparts.

1.2 The *Kannada* language

Kannada is one of the most well known Dravidian languages of India (Steever 2015) and is one among the twenty-two other official languages recognized by the constitution of India. Kannada is state language of Karnataka state and has history of more than 2000 years. There are more than sixty million Kannada speaking people in and out of Karnataka state. The Kannada language script is basically syllable-like structure and is called *Aksharas* which is evolved from *Kadamba script*. The language uses fifty two phonemic letters which are segregated into three groups viz., *Swaragalu*, the vowels, the *Vyanjanagalu*, the consonants and *Yogavaahakagalu*, the characters which are neither vowel nor consonant (Shridhara et al. 2013). Any Kannada word can be formulated using a meaningful combination of these Kannada *Aksharas* (Kannada alphabets). Further, it is possible to sub-divide syllable-like

Kannada alphabets into smaller consonant-vowel units which represents Phonemes of Kannada. The Kannada “Akshara” system basically consists of three different types of symbols. First, the consonants with an inherent vowel (/Ca/, the second type are consonants with other vowels (/CV/) and the third type are the consonant clusters followed by an vowel(/CCV/).

The details of the Kannada speech corpus used in the present study are discussed in Sect. 3.1.

2 Speech recognition system

A simple model of the basic speech recognition processes is given in Fig. 1. In general, automatic speech recognition (ASR) system involves several steps: viz. speech signal acquisition, pre-processing, cepstral analysis of speech and word recognizer. The first three stages constitute the front-end of ASR system.

The speech sub-word models are built using HMM and Viterbi decoding algorithm is used in recognizing words. The initial pre-processing step ensures good quality of the acquired input speech against recording levels and background noise. Speech analysis is carried out in cepstral domain to extract cepstral parameters. Mel frequency cepstral coefficients have been extensively used in speech recognition machines (Lawrence Rabiner 2012; Davis and Mermelstein 1980) and have proved to be good features in representing acoustic units. The main building block of speech recognition system is the pattern classifier. This has been accomplished by the well-known state of the art statistical model, Hidden Markov Model (HMM). The ASR system implementation basically has got two phases: viz., training phase and testing phase. In the training step, HMM based acoustical models are developed from the known training data set and during the testing phase, unknown speech samples are applied to evaluate the performance of the ASR system. Recognition accuracy is calculated based on the percentage of words correctly recognized.

2.1 Speech preprocessing

The speech signal preprocessing primarily involves pre-emphasis and windowing operations. Pre-emphasis filter (Deller Jr et al. 1993) is basically a high pass filter used to spectrally flatten the speech signal. After pre-emphasizing step, speech signal is segmented and sliced into overlapping time frames known as windowing or framing. This step is essential due to the fact that speech signal is basically non-stationary in nature. In choosing the window frame, there are two considerations viz., the type of window and the length of the window. Larger window length gives better spectral resolution while smaller window length will yield better time resolution (Lawrence Rabiner 2012). Typical window length of about 1–2 pitch periods is used by the researchers in the speech analysis (Rabiner 1989). In the present study, hamming window of 20 ms duration with 50 % overlapping has been used.

2.2 Cepstral analysis of speech

The feature analysis is the next important and critical stage in speech recognition system. The performance of the ASR greatly depends on the effectiveness of feature extraction algorithms used. Feature extraction stage seeks to provide compact encoding of the speech waveform. This encoding must minimize the information loss and provide a good representation of the acoustic model.

The present system considers MFCC as a feature set for the speech recognition processes. Human hearing is not equally sensitive at all frequency bands. The human ear perceivensness of frequency can be characterized by the Mel-Scale filter bank (Davis and Mermelstein 1980). The Mel-frequency cepstrum represents the short-term power spectrum of a sound, which is based on a linear cosine transform of a log power spectrum on the nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) collectively make up a Mel-frequency cepstrum. MFCCs are very commonly used in speech recognition systems and have proved to give good results (OShaughnessy 2008).

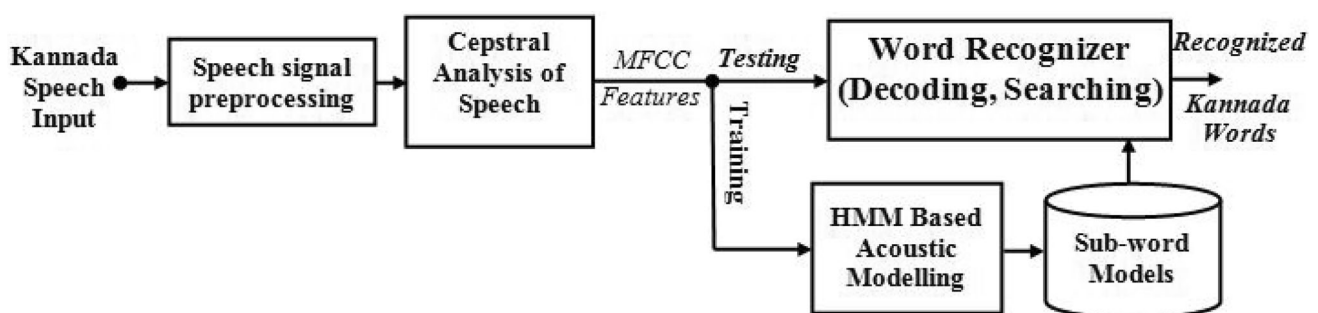


Fig. 1 Conceptual block diagram of the ASR system

In addition to MFCCs, an extra measure namely, logarithm of the signal energy has been appended for better representation. Also, it has been demonstrated that spectral transitions play an important role in human speech perception. Therefore it is desirable to consider and append the information related to time difference (or delta coefficients) and acceleration coefficients (second order time-derivatives). So, putting all together we have 12-MFCC parameters, one log energy value, 13 delta coefficients and 13 acceleration (delta-delta) coefficients. Hence the feature vector size becomes 39 in length.

2.3 HMM based acoustic modelling

In the modern ASR systems, HMM have become very common choice for acoustical modeling of speech sounds. The HMM can be assumed as a Markov process having hidden (or unobserved) states. They are double stochastic processes having one underlying state sequence that is hidden but may be estimated using a set of processes that produce observation sequence. The compact notation of HMM indicating the complete parameter set of the model is given by (1) :

$$\lambda = (A, B, \Pi) \quad (1)$$

where Π is initial state distribution vector, A is matrix of state transition probabilities and B is the set of the observation probability distribution in each state (Rabiner 1989; Nilsson 2005). In isolated word recognition (IWR) system, the goal is to find the closest matching word for a given word from the vocabulary size of say L words. Assuming that the observation sequence (MFCC feature vector in this case) be represented by (2):

$$O = \{o_1, o_2, \dots, o_M\} \quad (2)$$

Let the words to be recognized be represented by sub-word units according to (3) as follows:

$$W = \{sw_1, sw_2, \dots, sw_N\} \quad (3)$$

Then, the following Eq. (4) estimates the most likely word out of the L word vocabulary, for a given observation sequence O .

$$\hat{W} = \max_{W \in L} [P(W/O)] \quad (4)$$

Computing the probability of word W given the observation sequence, i. e. $P(W/O)$ is not practical and therefore Bayes rule is applied to make it feasible. So, by applying the Bayes rule on (4) and simplification gives (5).

$$\hat{W} = \max_{W \in L} [P(O/W)P(W)] \quad (5)$$

Equation (5) above, $P(W)$ is the prior probability of having the particular model, which produces the word to be

recognized. For a given word model W , the probability of having a particular observation sequence O is $P(O|W)$ and can be obtained by using HMM (Lawrence Rabiner 2012). In the Eq. (5), the probability of observation sequence $P(O)$ has been ignored because, this probability is constant when attempting to maximize over the possible word models. The procedure followed in the isolated word recognition can be summarized into following two steps:

1. Build individual HMM with same number of states for each sub-word units in the vocabulary. This step is basically the training phase, where HMM parameters are estimated using Baum-Welch re-estimation technique to obtain the sub-word model.
2. The second step is the testing phase, where trained HMMs are used to identify each unknown word in the testing data set. The recognition phase is a decoding problem, where the best state sequence for given sub-word model is computed using Viterbi algorithm to obtain the best matching word.

Finally, the performance of the system is evaluated based on the word error rate (WER) or system accuracy is computed in terms of percentage of words recognized.

3 ASR system implementation using HTK

The implementation of speech recognition system for Kannada words has been carried out using Hidden Markov Tool Kit (Young et al. 1997) (HTK version 3.4.1) in the Linux platform. The implementation involves mainly data preparation, data coding, acoustic modeling using HMM and finally evaluating the performance of the system.

Data preparation is the stage where speech signal is acquired in a controlled environment and pre-processed. Also, the vocabulary of the system is defined using a pronunciation dictionary which contains labels and phones corresponding each of the words. In the data coding (speech analysis) step, speech waveforms are parameterized to obtain the sequence of feature vectors. All the necessary specifications such as frame rate, frame size, frame type and features required are to be set during speech analysis. Next, the HMMs for all the sub-word units are defined and sequences of feature vectors are used to re-estimate and generate the HMMs. Further, the set of test database are applied on the recognizer to evaluate the system performance.

3.1 Speech corpus

The speech signal recording has been carried out by using a good quality audio recording device (Linear PCM

recorder) in a controlled environment (minimum noise, sound proof room) and stored in wave file format. Speech database used in this study has been developed by recording the regional Kannada broadcast news corpus uttered by six different speakers. The recordings consists of three male and three female speakers with recording duration of about 10 min each (total of about 60 min of speech data). An example of Kannada news text corpus used for creating database is shown in Fig. 2. The chosen Kannada news text corpus consists of 5468 Kannada words in which 1500 words are unique (distinct words without any repetitions). Therefore the dictionary size of the system becomes equal to 1500 words. The word segmentation has been carried out manually. In order to obtain the better performance accuracy in an ASR system a large amount of speech recordings are required. But due to limitations of speech database size, Lachenbruchs holdout procedure has been employed (Johnson et al. 1992). Accordingly, the entire speech database has been sub-divided into four groups (namely, Group-1, Group-2, Group-3 and Group-4). The first three groups consists of utterances of 1056 words each obtained from the recordings of first four speakers (two male and two female). The fourth group consists of utterances of 2300 words obtained from the recordings of fifth and sixth speakers (one male and one female). In the present study, the first three groups are used for both training and testing whereas, the fourth group is exclusively used for testing. This grouping of database has been done arbitrarily and are purely unbiased.

The phone set for Kannada language is represented by one or two English alphabets as given in Table 1. Phone-level and syllable-level acoustical models are built to evaluate the recognition performances. Total number of phones in a given language is fixed whereas the number of syllables increases with the increase in vocabulary size.

Table 1 Phones in Kannada language

Label	Kannada phoneme	Label	Kannada phoneme	Label	Kannada phoneme	Label	Kannada phoneme
a	ಅ	au	ಔ	T	ಟ್	bh	ಭ್
A	ಆ	M	ಠ(Ex.ಅಠ)	Th	ಥ್	m	ಮ್
i	ಇ	H	ಃ(Ex. ಅಃ)	D	ಡ್	y	ಯ್
l	ಈ	k	ಕ್	Dh	ಢ್	r	ರ್
u	ಉ	kh	ಖ್	N	ಢ್	l	ಲ್
U	ಊ	g	ಗ್	t	ತ್	v	ವ್
ru	ಋ	gh	ಘ್	th	ಠ್	sh	ಶ್
rU	ೠ	ng	ಙ್	d	ಡ್	S	ಷ್
e	ಎ	c	ಚ್	dh	ಢ್	s	ಸ್
E	ಏ	ch	ಛ್	n	ನ್	h	ಹ್
ai	ಐ	j	ಜ್	p	ಪ್	L	ಲ್
o	ಒ	jh	ಝ್	ph	ಫ್	kSh	ಕ್ಷ್
O	ಓ	ny	ಞ್	b	ಬ್	jn	ಜ್ಞ್

Two pronunciation dictionaries namely, phone-level and syllable-level dictionaries are constructed for the available words in the database. In a phone-level pronunciation dictionary, sequence of phones represents a particular word, where as in syllable-level, sequence of syllables represents a word. There are about 500 syllables for the vocabulary size of 1500 Kannada words. Syllables are basically combination of more than one phone and are represented by the characters (orthographic structure) of the Kannada language. A few examples for representation of Kannada words as sequence of syllables and phones are given in Table 2.

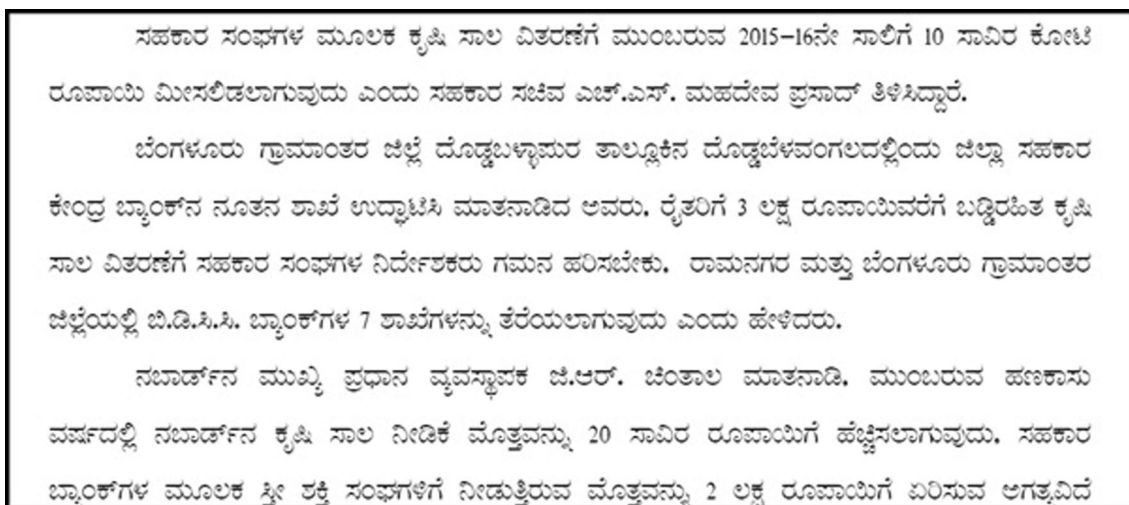


Fig. 2 Example of news text corpus used in creating speech database

Table 2 Examples of phone and syllable sequences for Kannada words

Kannada word	English meaning	Sequence 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 ka
ಸಾವಿರ (savira)	Thousand	sA vi ra	s A v i r a
ಅವರು (avaru)	They	A varu	a v a r u
ನೂತನ (noothana)	New	nU ta na	n U t a n a

3.2 Feature analysis

Figure 3 shows the block diagram of speech coding step in HTK. Speech waveforms (wave files) are parameterized into MFCC feature vectors as per the specifications (using HCopy command) as shown in Table 3. Each feature vector consists of 12-MFCC values, 1-log energy value, 13-delta coefficients (first derivative values) and 13-delta-delta coefficients (second derivative values). Hence the feature vector size becomes 39.

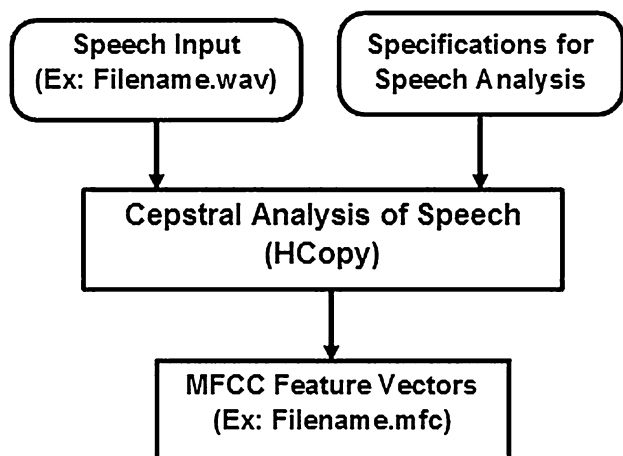


Fig. 3 Speech coding steps using HTK tool

Table 3 Speech coding specifications

Parameters	Specifications
Sampling rate	16,000 Hz
File format	16-bit, mono, wave file
Window length	20 ms
Window used	Hamming
Window overlapping	50 %
Pre-emphasis coefficient	0.97
Features used	MFCC, log energy value Δ MFCC and $\Delta\Delta$ MFCC
Feature vector length	39
Number of filter-banks	26
Number of MFCC coefficients	12

Analysis window length is chosen to be 20 ms with 50 % overlapping between the adjacent two frames. So feature vector is generated for every 10 ms duration. Here speech parameterization is done for entire data-base including the training and testing samples. The sequence of feature vectors of training samples are used to build the HMMs, whereas the one pertaining to testing samples are used to evaluate the models.

3.3 Generating HMM

The acoustical modeling of sub-word units are carried out with the help of HMMs. Mono-phone and Triphone HMM models for the Kannada words are explored in the present study. Both the phone-level and syllable-level sub-word acoustical models have been considered in the design of monophone as well as triphone HMM models. Each sub-word model is designed using a 5-state HMM with first and last states (states S1 and S5) being non-emitting states. These are required to model the word boundaries (silence states). The middle three states (states S2, S3 and S4) represent the sub-word unit and these states are assumed to be of Gaussian type. Figure 4 shows the structure of a 5-state HMM design. The phone-level modeling uses all 52 Kannada phones whereas syllable-level model requires 500 syllables in the present study. The meaningful sequence of phones or syllables forms a particular word as given in Table 2.

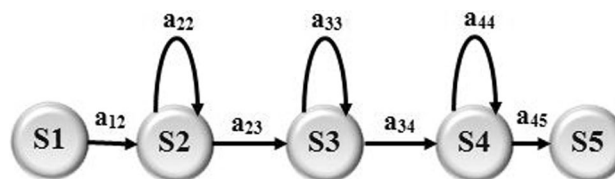


Fig. 4 Structure of a five-state HMM

Initially, context independent mono-phone HMM models are generated by using MFCC feature vectors obtained from the training speech database. Estimation and re-estimation of sub-word models are carried out with help of HRest function of HTK tool. The HRest function adapts Baum-Welch algorithm for model re-estimation of HMM (training). Mono-phone HMM models are further extended to obtain context dependent triphone models. The conversion from monophone into triphone has two steps. The first step is to convert the monophone transcriptions into triphone transcription and then the triphone models are re-estimated. In the second step, tying of equivalence acoustic states is carried out to confirm the proper estimation of all state distributions (Nilsson 2005).

After building the HMMs (training step); performance of the recognizer has been evaluated on the testing data set. For all the test samples, Viterbi decoding algorithm (HVite command in HTK) has been used to find the best matching word from the dictionary. Finally, result analysis has been done (using HResults command) by observing the percentage of word recognition accuracy. The accuracy is the ratio of number of word recognized to the total number of words present.

4 Results and discussions

Recognition performance of an isolated word recognition system is quantified by using the word recognition accuracy. The Table 4 shows the details of results obtained for the isolated Kannada word recognition system implemented in the present study. The word recognition accuracy is calculated according to the following expression (6);

$$\text{Percentage of Accuracy} = \frac{N - S}{N} \times 100\% \tag{6}$$

where N is the total number of words in the test set and S represents number of words substituted (or number of words misclassified). An average accuracy of 74.35 % has been reported for a triphone based recognition system. The plot in Fig. 5 shows recognition performances of phone-level and syllable-level acoustic models for various testing

groups. On comparing the phone-level and syllable-level mono-phone HMM performances, we can clearly observe that phone-level model performs better than the syllable-level based system. In our previous study (Ananthakrishna et al. 2015) based on small vocabulary systems it has been reported that, syllable-based HMM performance was better than that of the phone-level counterpart. In small vocabulary systems, number of syllables is comparable to the number of phones. But, as the dictionary size increases, the number of syllables required to be modelled also increases significantly. This has the adverse effect on the performance of the recognition system. Further, the amount of training that each syllable models receive is significantly less than that of phone models for the given training data set.

Monophone models (phone-level or syllable-level acoustic models) are basically context independent models. In a continuous speech, phones are influenced by the presence of its adjacent phones. In order to capture the phone-to-phone variations, context dependent triphone models are considered. The comparative word recognition accuracy of monophone and triphone HMM systems is plotted in Fig. 6. We can observe a significant improvement in the results of triphone models against that of the monophone models. Hence context dependent acoustical models are found to be good for *Kannada* word recognition system in this study. In fact, these results are in line with the performances of conventional speech recognition systems.

The HMM states in monophone (both phone-level and syllable-level) and triphone acoustical models assume simple Gaussian type (single Gaussian distribution). But with the increase in the number of Gaussian mixtures of each state, the model better represents the acoustical sub-word unit. Therefore in the present study, recognition accuracy of monophone system has been evaluated by varying the number of Gaussian mixtures in each state of HMM and the results are as shown in Table 5. It can be seen that monophone system performance increases with the increase in the number of Gaussian mixture components. With the increase in mixture components, monophone

Table 4 Results: word recognition accuracy (%)

Trial no.	Testing data group	Mono-phone HMM		Tri-phone-HMM
		Phone-level Model	Syllable-level Model	
1	Group-1	58.38	41.62	77.89
2	Group-2	62	35.73	75.99
3	Group-3	62.59	36.46	78.98
	Average	60.20	36.39	74.35
4	Group-4	57.82	31.74	64.54

Fig. 5 Recognition accuracy of phone-level versus syllable-level model

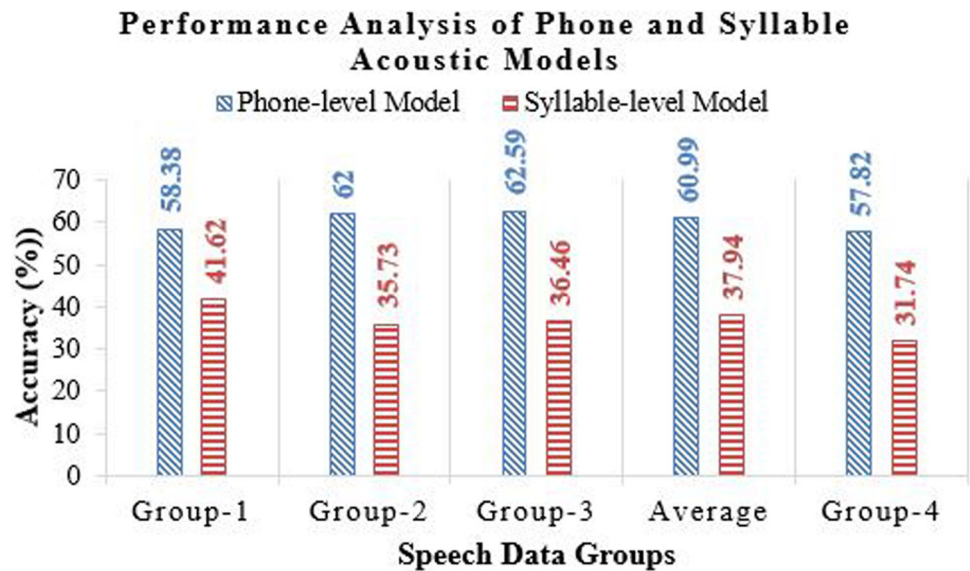


Fig. 6 Recognition accuracy of mono-phone versus tri-phone model

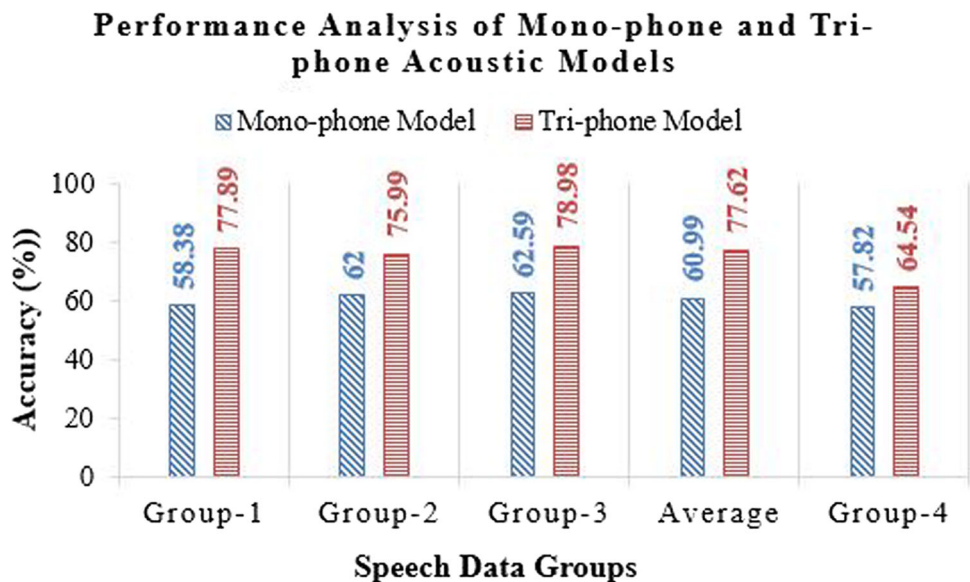


Table 5 Recognition accuracy of mono-phone models for different Gaussian mixtures

No. of Gaussian mixtures	Speech Data Groups				
	Group1	Group2	Group3	Avg.	Group4
2	60.86	64.08	64.02	62.99	61.13
4	62.86	67.11	65.15	65.04	62.79
8	64	68.34	66.67	66.34	64.31
16	66.38	69.94	67.14	67.82	64.84
32	70.38	71.27	67.33	69.66	66.49
64	72.19	71.83	69.6	71.21	67.1
128	72.86	72.87	71.02	72.25	68.54

system performance almost attains that of triphone system. In a GMM, each HMM state is a multivariate Gaussian density function, which better represents the sub-word

acoustic unit and hence gives better results. But as the number of mixture increases the computation complexity also increases exponentially. For each speech input frame,

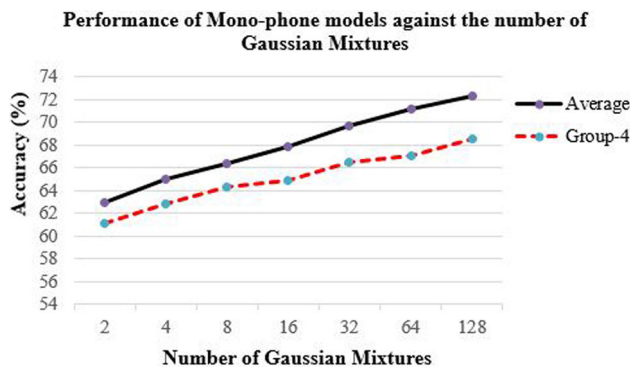


Fig. 7 Performance of mono-phone models for different gaussian mixtures

the output likelihood should be evaluated against each active HMM state. Therefore, the state likelihood estimation is computationally very intensive and hence this may not be useful for real-time applications.

Figure 7 shows the plot of performance of monophone models against number of Gaussian mixtures. The overall average recognition accuracy of first three data groups (Average of Group-1, Group-2, and Group-3 results) is about 2–5 % more than that of Group-4. Similar observations can be made in plots of Figs. 5 and 6. It can be noted that the Group-4 consists of speech data pertaining to two “new” speakers (1 male and 1 female speaker) which are not used during training phase. But this is not the case in the remaining three data groups. The *holdout* procedure has been followed for the first three data groups, where one of the data group is used for testing while the remaining two are used for training. And each of the Groups-1, 2 and 3 consists of recordings from all the four speakers. Also, Group-4 has more number of “unseen” words by the recognizer and hence this accounts for the poor performance of Group-4 test data.

5 Conclusion

In this study, HMM based isolated speech recognition system for the Kannada words has been implemented using HTK software tool. This implementation involves development of the Kannada news database and building the pronunciation dictionaries (both phone-level and syllable-level) for the chosen vocabulary. The ASR system performance has been evaluated for monophone, mono-syllable and triphone acoustical models. Best average word recognition accuracy of 74.35 % has been obtained for triphone based system. It is observed that the monophone based system performs better than mono-syllable systems for the chosen medium sized vocabulary of about 1500 words. It would be therefore preferable to choose phone-level models

against syllable-level models for the large vocabulary systems. System performance can be further improved by adapting context dependent acoustical models. Hence, triphone based acoustical models are better than the monophone based systems. Further, the work could be extended by increasing the number of Gaussian mixtures in triphone models. Also, Kannada language specific features may be explored to improve the system performance.

References

- Aggarwal, R., & 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.
- Aggarwal, R., & Dave, M. (2012). Integration of multiple acoustic and language models for improved Hindi speech recognition system. *International Journal of Speech Technology*, 15(2), 165–180.
- Ananthakrishna, T., Maithri, M., & Shama, K. (2015). Kannada word recognition system using HTK. In: *2015 annual IEEE India conference (INDICON)* (pp. 1–5). IEEE.
- 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, 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. B., & 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., Jr., Proakis, J. G., & Hansen, J. H. (1993). *Discrete time processing of speech signals*. Upper Saddle River: Prentice Hall PTR.
- Hassan, F., Kotwal, M. R. A., Muhammad, G., & Huda, M. N. (2011). MLN-based bangla ASR using context sensitive triphone HMM. *International Journal of Speech Technology*, 14(3), 183–191.
- Hegde, S., Achary, K., & Shetty, S. (2012). Isolated word recognition for kannada language using support vector machine. In: *Wireless networks and computational intelligence* (pp. 262–269). Berlin: Springer.
- Hegde, S., Achary, 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.
- Hemakumar, G., & Punitha, P. (2014b). Speaker dependent continuous Kannada speech recognition using HMM. In: *2014 international conference on intelligent computing applications (ICICA)* (pp. 402–405). IEEE.
- Hemakumar, G., & Punitha, P. (2014a). Automatic segmentation of Kannada speech signal into syllables and sub-words: Noised and noiseless signals. *International Journal of Scientific & Engineering Research*, 5(1), 1707–1711.
- Johnson, R. A., Wichern, D. W., et al. (1992). *Applied multivariate statistical analysis* (Vol. 4). Englewood Cliffs, NJ: Prentice Hall.
- Kumar, K., & Aggarwal, R. K. (2012). A Hindi speech recognition system for connected words using HTK. *International Journal of Computational Systems Engineering*, 1(1), 25–32.

- Lakshmi, A., & Murthy, H. A. (2006). A syllable based continuous speech recognizer for Tamil. In: *INTERSPEECH*.
- Mannepalli, K., Sastry, P. N., & Suman, M. (2016). MFCC-GMM based accent recognition system for Telugu speech signals. *International Journal of Speech Technology*, 19(1), 1–7.
- Mishra, A., Chandra, M., Biswas, A., & Sharan, S. (2011). Robust features for connected Hindi digits recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 4(2), 79–90.
- Muralikrishna, H., Ananthakrishna, T., Shama, K. (2013). HMM based isolated Kannada digit recognition system using MFCC. In: *2013 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 730–733). IEEE.
- Neti, C., Rajput, N., Verma, A. (2002). A large vocabulary continuous speech recognition system for Hindi. In *Proceeding of works multimedia signal processing* (pp. 475–481).
- Nilsson, M. (2005). First order hidden markov model: Theory and implementation issues. Research Report, February 2005, Department of Signal Processing, Blekinge Institute of Technology.
- OShaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10), 2965–2979.
- Panda, S. P., & Nayak, A. K. (2015). Automatic speech segmentation in syllable centric speech recognition system. *International Journal of Speech Technology*, 19(1), 1–10.
- 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., & Juang, B. H. (2012). *Fundamentals of speech recognition*. Upper Saddle River: Prentice Hall.
- Radha, V., et al. (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*, 4(6), 2223–2229.
- Shridhara, M., Banahatti, B. K., Narthan, L., Karjigi, V., & Kumaraswamy, R. (2013). Development of Kannada speech corpus for prosodically guided phonetic search engine. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)* (pp. 1–6). IEEE.
- Steever, S. B. (2015). *The Dravidian languages*. London: Routledge Publications.
- Sunitha, K., Kalyani, N., et al. (2012). Isolated word recognition using morph-knowledge for Telugu language. *International Journal of Computer Applications*, 38(12), 47–54.
- Thangarajan, R., Natarajan, A., & 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., et al. (1997). *The HTK book* (Vol. 2). Cambridge: Entropic Cambridge Research Laboratory.