Development of Speech corpora for different Speech Recognition tasks in Malayalam language

Cini kurian
Al-Ameen College, Edathala, Aluva

***Abstract*—Speech corpus is the backbone of an Automatic speech Recognition system. This paper presents the development of speech corpora for different speech recognition tasks in Malayalam language. Pronunciation dictionary and Transcription file which are the other two essential resources for building a speech recognizer are also being created. Speech recognition performance of different speech recognition tasks are being presented. Speech corpus of about 18 hours have been collected for different speech recognition tasks.**

Keywords— Speech Recognition, corpus development, Malayalam

I. INTRODUCTION

One of the main challenges faced by speech scientist is the unavailability of the three important resources. The prime and most importantly, speech corpus (Speech Database), pronunciation dictionary and transcription file. Very fewer efforts have been made in Indian languages to make these resources available to public compared to English. Creation of these resources is time consuming, boredom and needs so much man power. Creating a well defined pronunciation dictionary needs through knowledge from phonetics, phonological rules, syntactic and semantic structure of the language.

It is necessary to have databases which comprises of appropriate sentences spoken by the typical users in realistic acoustic environment. Speech databases can be divided into two groups: (i) a database of speech normally spoken in a specific task domain. In this case, small amount of speech is sufficient to achieve acceptable recognition accuracy. (ii) a general purpose speech database that is not tuned to a particular task domain but consists of general text and hence can be used for recognition of any sentence in that language. The problem with most speech recognition systems is insufficient training data containing speech variations (spontaneous speech) caused by speaker variances (cover large number of speakers). To overcome these problems, a large vocabulary speech database is required to build a robust recognizer. The purpose of selection of phonetically-rich sentences is to provide a good coverage of pairs of phones in the sentence. The current work also aims at the development of databases for Malayalam speech recognition that will facilitate for acoustic phonetic studies, training and testing of automatic speech recognition systems. It is anticipated that the availability of this speech corpus would also stimulate the basic research in Malayalam acoustic-phonetics and phonology.In this paper three sets of databases have been created (task specific databases , a general purpose database and a specially designed database for unique phoneme analysis) . The task specific database includes three domain based databases i.e isolated digit speech database, connected digit speech database and continuous speech database. The general purpose database includes a set of phoneme class wise speech database. Database designed for unique phoneme analysis includes, specially designed 32 minimal pair of words as well as a set of words which include unique phonemes in any word positions.

Section 2 discusses phonetic chart of Malayalam language and in section 3 the text corpus that has been prepared is detailed. In section 4 the method of speech data collection is being elaborated. The phoneme list prepared for the work is explained in section 5 followed by the creation of pronunciation dictionary in section 6. In section 7 the format and model of the transcription files being prepared is explained. Section 8 gives the results of various speech recognition tasks.

2. PHONETIC CHART

Malayalam has 52 consonant phonemes, encompassing 7 places of articulation and 6 manners of articulation, as shown in Table 1 below. In terms of manner of articulation, plosives are the most complicated, for they demonstrate a six-way distinction in labials, dentals, alveolar, retroflex, palatals, velars and glottal [1]. A labial plosive, for example, is either voiceless or voiced. Within voiceless labial plosives, a further distinction is made between aspirated and un-aspirated ones whereas for voiced labial plosives the distinction is between modal-voiced and breathy-voiced ones. In terms of place of articulation, retroflex are the most complex because they involve all manners of articulation except for semi vowels [2]. Phonetic chart as presented by Kumari, 1972 [3] for Malayalam language is given in table 1 and the same has been referred for this paper.

For all speech sounds, the basic source of power is the respiratory system pushing air out of the lungs. Sounds produced when the vocal cords are vibrating are said to be voiced, where the sound produced when the vocal cords are apart are said to be voiceless [4]. The shape and size of the vocal tract is a very important factor in the production of speech. The parts of the vocal tract such as the tongue and the

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 229–236, Trivandrum, India. December 2015.

lips that can be used to form sounds are called articulators (fig .1). The movements of the tongue and lips interacting with the roof of the mouth (palate) and the pharynx are part of the articulatory process [5]

Table 1 : Phonetic chart of Malayalam

		Labial		Dental		Alveolar		Retroflex		Palatal		Velar		Glottal
		voiced	unvoiced	voiced	unvoiced	Voiced	Unvoiced	voiced	unvoiced	voiced	unvoiced	voiced	unvoiced	
Stop / plosive	un aspirated	പ p	ബ b	ത t	ദ d	റ t̲		ട ṭ	ഡ ḍ	ച c	ജ j	ക k	ഗ g	
	aspirated	ഫ ph	ഭ bh	ഥ th	ധ dh			ഠ ṭh	ഢ ḍh	ഛ ch	ഝ jh	ഖ kh	ഘ gh	
Nasals		മ m		ന n		ന n̲		ണ ṇ		ഞ ñ		ങ ṅ		
Fricative		ഫ f		സ s				ഷ ṣ		ശ ś				ഹ h
Lateral						ല l		ള ḷ		ഴ l				
rhotic						ര r		റ r						

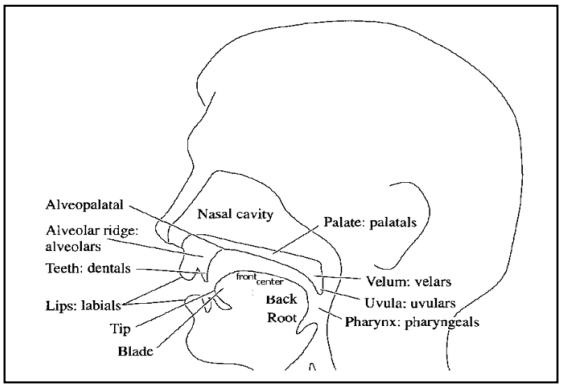


Figure 1: Place of articulation

3. Text Corpus

The first step followed in creating the speech database for automatic speech recognizers is the generation of optimal set of textual sentences to be recorded from the native speakers of the Malayalam language.

The following are the different set of text corpus collected for different tasks.

3.1 Isolated Digit Recognition Task

Digits from 0-9

3.2 Connected Digit Recognition Task

Text corpus consists of specially designed, twenty, 7 digit numbers so that maximum combinations of numbers as well as co-articulation of the numbers have been taken into account. It is assumed that speakers will be comfortable in reading 7-digit numbers. Accordingly, a sets of 7-digit numbers were generated, each set containing 20 numbers capturing all distinct word pairs. These sets were generated using different methods of generating word pairs. One such set of 20 numbers is shown in Table 2.

Table 2: An example of a list of twenty 7-digit numbers read by speakers

0098765	4159306	4725836	3567801
5432101	6927918	6828162	1345566
1975312	8908634	2612371	6778899
2964203	4074851	1460450	9011217
3952494	1733844	0570223	7913579

3.3 Continuous Speech Recognition task

For speaker independent speech recognition, speech data needs to be collected from a large number of speakers. It is not practical to ask speakers to speak/read a lot of sentences that contain all phonemes (in various phonetic contexts) of the language. Hence, it is desirable to construct sets of sentences that are phonetically rich. This construction is a laborious task. Phonetically rich sentences can be selected from a large set of text. Traditional sources of text data are books, magazines and periodicals. However a textual data is needed in electronic form so that it can be processed by a computer. Hence there are two choices: (a) Manually type in the printed data from articles, periodical, magazines etc., and store it in electronic form. (b) Use available online sources of text data. Example of such sources is articles on web and online news papers. There

are several online newspapers that provide content in Malayalam language. Each online content uses its own grapheme encoding scheme to display Malayalam text. To collect the text data online Malayalam news papers being used. In order to compute the phonetic richness of sentences and select sentences, the Malayalam text (and the corresponding phoneme sequence) has to be represented using Roman symbols. However, such a grapheme to phoneme conversion programme for the different fonts was not available. Also, information about the coding scheme was not readily available. Hence a grapheme-to-phoneme (G2P) program was written to take care of most conventions of the font. A tools named Corpuscrt [6] from CMU (Carnie Melon University) is used to select the maximum phonetically rich sentences. Hence the selected text includes about 202 sentences, comprising of about 1600 words.

3.4 Analysis of Unique Phoneme Features for speech recognition

For the analysis of unique phoneme features of Malayalam language, the phonemes Alvelor plosive ഀ (t't'a) , Alvelor rhotic ഠ (ra) , Retroflex lateral ഩ **(l'a)** ,Palatel lateral ഴ (zha) and Dental Nasal ഩ (n1a) have been selected. For this analysis, two set of words have been compiled.

i. The first set contains a word-list constituting each of these sounds in all permissible word-positions. In this category, a total of 340 words have been compiled. The following are (Table 3) the total number of section wise, categorization of words with these phonemes that have been collected for the study.

Table 3: List of Words with different categories of phonemes

Category	Total numbers
Alvelor plosive	27
Alvelor rhotic	48
Retroflex lateral	37
Palatel lateral	42
Dental Nasal	40
retroflex rhotic	49
alveolar lateral	44
alveolar nasal	53

ii. Since /la/ and /l̪'a/, /ra/ and /r'a/ are two pairs of contrastive phonemes and /zha/ contrasts in some instances with one or both of the laterals or in other instances with one or both of the rhotics, the data sets are designed to include minimal pairs. Accordingly 32 minimal pairs have been compiled (total of 64 words). In short a total of 404 words were designed for unique phoneme study.

3.5 Phoneme class wise speech recognition task

Five phoneme classes of words have been chosen for this task. The different phoneme classes are stop, lateral, fricative, nasals and rhotic. Maximum words have been included such that it should contain all phonemes in all word positions (start, middle and end).

Nasal class category contain 74 words which include labial, dental, alveolar and retroflex nasals. Category wise number of words is shown in table 4.

Table 4: List of words with Nasal class phonemes

Category	Total numbers
Labial nasals	20
Dental nasals	20
Alveolar nasals	14
Retroflex nasals	20

i. Lateral class words include a set of 54 words carefully designed which includes palatal, retroflex and alveolar lateral as detailed in table 5.

Table 5: List of words with lateral class phonemes

Category	Total numbers
Palatel lateral	15
Retroflex lateral	15
Alveolar lateral	24

ii. ii. Fricative class words include a set of 70 words which includes dental, retroflex and palatal fricatives. Table 6 list the number of words of each type.

Table 6 : List of words with fricative class phonemes

Category	Total numbers
Dental fricative	23
Retroflex fricative	15
Palatel fricative	17
Glotal fricaitive	15

iii. Plosive/stop class words includes a total of 205 words in different categories such as labial voiced stop, labial unvoiced stop, dental voiced stop, dental unvoiced stop, velar voiced stop, velar unvoiced stop, retroflex unvoiced stop, palatal voice stop and palatal unvoiced stop. Table 7 lists the number of words in each category.

Table 7 : List of words with plosive/stop class phonemes

Category	Total numbers
Labial voiced stop	26
Labial unvoiced stop	20
Dental voiced stop	20
Dental unvoiced stop	20
Retroflex voiced stop	21

Retroflex unvoiced stop	15
Palatel voiced stop	20
Palatel unvoiced stop	22
Velar voiced stop	23
Velar unvoiced stop	19

iv. Rhotic class words includes 43 words which include both rhotics (alveolar rhotic and retroflex rhotics) as detailed in table 8.

Table 8: List of words with rhotic class phonemes

Category	Total numbers
Alvelor rhotic	23
Retroflex rhotic	20

4. Speech data collection

Speech data for all the recognition tasks is collected from the age group of 20 to 45 years keeping almost equal male and female ratio. Speakers were requested to read the word /sentences in a normal reading manner. Speech data is collected in normal office environment using a microphone with 1600 frequency. Mistakes made while recording have been corrected by re-recording or by making the corresponding changes in the transcription file. For all the tasks data have been collected from 25 speakers (13 female and 12 male speakers). The task wise list of speech corpus is detailed below.

- ***Isolated Digit recognition task*** - Digits zero to nine is uttered separately by speakers. Hence Size of the speech corpus for this task is 250 words.
- ***Connected digit recognition task*** - Text data as detailed in section 3.2 is read by the 25 speakers in normal reading manner. Size of the corpus is 500 words.
- ***Continuous speech recognition task***- The 202 continuous sentences which are selected as described above were read by the speakers. Hence the size of the continuous speech corpus is 40000 words.
- ***Unique, phoneme study*** - The specially designed 404 words were read 25 speakers thereby enhancing the speech corpus by 10100 words.
- ***Phoneme class wise recognition task*** - For this category of speech corpus the 446 words as mentioned in section 3.5 were read by 25 speakers. Hence the size of speech corpus in this category is 11150 words.

These details are given in table 9 below.

Table 9: Speech corpus collection

Recognition task	Total text size in words	Number of speakers	Total size of speech corpus in words	Speech corpus in hrs
Isolated digit	10	male : 12 female :13	250	0.16
connected digit	20	male : 12 female :13	500	0.63
continuous speech recogntion task	1600	male : 12 female :13	40000	11.67
Analysis of				

5. Creation of Phoneme List

A phoneme is the basic unit of recognition. Therefore the preparation of a phoneme list is a vital step in creating a pronunciation dictionary. Each phone to be denoted by a set of phonetic notation rather than a single notation, which will be decided by the acoustic property of the phones. All unique phonemes of Malayalam language have been identified and phonetic notation has been assigned according to its phonetic properties. Table 10 lists all the phonemes used in this work along with its phonetic notation.

Table 10: Phoneme list

Malayalam	Phonetic Notation	Malayalam	Phonetic Notation	Malayalam	Phonetic Notation
അ	a	ച്	clch ch	മ്	m
ആ	aa	ഛ്	clch chch	യ്	y
ഇ	i	ജ്	vbj j	ര്	r
ഈ	ii	ഝ്	clch chch	ല്	l
ഉ	u	ഞ്	nj'	വ്	v
ഊ	uu	ട്	clt' t'	ശ്	sh
ഏ	e	ഠ്	clt' t'h	ഷ്	s'h
ഏ	e'	ഡ്	vbd' d'	സ്	s
ഐ	ai	ഢ്	clt' t'h	സ്	s1
ഒ	o	ണ്	n'	ഹ്	h
ഓ	o'	ത്	clt t	ള്	l'
ഔ	au	ഥ്	clt th	റ്	r'
ം	m	ദ്	vbd d	ഴ്	z
ഃ	-	ധ്	clt th	റ്റ്	clr1 r1
്	u'	ന്	n (Dental nasal)	ഫ്	ph1 (fan)
ക	clk k	ന്	n1(alveolar nasal	ള്	l'
ഖ്	clk kh	പ്	clp p	ണ്	n'
ഗ്	vbg g	ഫ്	clp ph	ന്	n1
ഘ്	clk kh	ബ്	vbb b	ര്	r'
ങ്	ng'	ഭ്	clp ph	ല്	l

6. Pronunciation Dictionary (PD)

In pronunciation dictionary all words in the training data to be mapped onto the acoustic units which are defined in the phone list. Theoretically, creation of phonetic dictionary is just a mapping of grapheme to phoneme. But this alone would not work especially for a language like Malayalam as many phonemes pronounced differently in different contexts. For example, ഫ (ph'a) pronounced differently in ഫലം (/ph'alam/-fruit) and ഫാൻ (/ph'aan'/ - fan) and ന (n1a and

na - Nasal dental and Nasal alveolar) is pronounced differently even though the grapheme notation is same (eg. നനക്കുക(/n1anaykkuka/- watering). Hence for creating pronunciation dictionary, initially mapping have been completed for all grapheme into the corresponding phoneme units. Then some phonological rules have been applied manually and edited the dictionary. Multiple pronunciations are also incorporated in the dictionary.

The format of pronunciation dictionary for isolated digit recognition is in table 11 and that of continuous speech recognition task is under table 12. The left segment represent orthographic transcription and right segment shows its actual pronunciation is taken from the training speech corpus. The dictionary creation to be done with utmost care and precision as the duplication of phone for different sounds will confuse the trainer and the model so created will give false information to the recognizer. The dictionary must have all alternate pronunciations marked with parenthesized serial numbers starting from (2) for the second pronunciation. The marker (1) is omitted for the first pronunciation. Thus pronunciation dictionary of 2480 words of Malayalam language have been prepared for different tasks.

The following are the size of pronunciation dictionary for different tasks/analysis

i. Digit recognition task - 10 words
ii. Connected word recognition tasks - 20 words
iii. Continuous speech recognition task - 1600 words
iv. Unique phoneme analysis - 404 words
v. Phoneme class wise recognition task - 446 words

Table 11: P.D for isolated digit recognition task

puujyam	clp p uu j y a m
onnu'	o n3 u'
raNtu'	r a n: vbd: d:
muunu'	m uu n3 u'
naalu'	n3 aa l u'
anchu'	a nj clc u'
aar'u'	aa r' u'
eezu'	ee zh u'
ettu'	e clt: t: u'
ompatu'	o m clp p a clt t u'
ompatu (2)	o n clp p a clt t u'

Table 12: Format of PD for continuous speech recognition task

```
aadyapaadattilum'   aa vbd d y a clpp aa vbd d a clt  tt i l u m
aago'l'avipan'iyil_(2)        aa vbg g o' l' a v i clp p a n' i y i l
aago'l'avipan'iyil_                    aa vbg g o' l' a v i clp p a n' i y i l
aakaashattu'ninnul'l'a        aa clk k aa sh a clt tt u' n i nn u l'l' a
aakar_s'hakamaayad'isainukal'il_        aa clk k a r' s'h a clk k a m aa y a vbd' d' i s ai n1 u clk k a l' i l
aalappuza     aa l a clp pp u zh a
aam'bulan_su'                  aa m vbb b u l a n s u'
aam'bulan_su'kit't'iyillenna           aa m vbb b u l a n s u' clk k i clt'l t'l i y i ll e n1n1 a
aandhraprade'shu'             aa n clt th r' a clp p r' a vbd d e' sh u'
aapuro'gatiyum'               aa clp p u r o' vbg g a clt t i y u m
aaram'bhichchatu'             aa r a m clp ph i clch chch a clt t u'
aaram'bhikkum'                aa r a m clp ph i clk kk u m
aaram'bichchu                 aa r a m vbb b i clch chch u
```

Transcription file

The transcription file contains the sequence of words transcribed orthographically, and non-speech sounds, written exactly as they occurred in the training speech, followed by a tag which can be used to associate this sequence with the corresponding training speech data. The acoustic speech file is transcribed into its corresponding orthographic representation. The transcription file is to be prepared for every training speech data after closely examining/hearing the wave file. Hence the transcription process is done manually considering even silence, noise or a breath. This is a herculean task and has to be done very carefully, since a minute error in the transcription file will mislead the recognizer and will lead into misclassification of training data. Hence transcription files prepared for a total of 62000 speaker utterances (wave file). Each task has a separate transcription file which consisting of transcriptions for each speaker utterance. Table 13 is a sample of transcription file created for continuous speech recognition task.

Table 13 : A sample of transcription file for continuous speech recognition task

```
<s> at'isthaanavilayil_ anj'chushatamaanam'maar_jin_ nalkan'amennaavashyappet't'ukon't'u' patimuunninu' pet'ro'l'_panpukal'_ at'achchit't'upratis'he'dhikkuvaan_ vitaran'akkaar_ tiirumaanichchu </s>
<s> atinit'e vivaadamaayasid'i pibiyut'enir_de'shaprakaaram' nir_michchataan'ennu nir_maataavum' sam'vidhaayakanum' po'liisino't'u sammatichchit't'un't'u' </s>
<s> ate'samayam' prashnattil_ vaadam'ke'l'_kkunnatu' teranj'nj'et'uppukammiis'han_ naal'atte'kkumaar'r'i </s>
<s> itu vivaadamaayappo'l'_ supriim'ko't'ati no't't'iisayachchu </s>
<s> intyan_ aayur_ve'davyavasaayam' naalaayirattiirunnuur'r'ianj'chu' ko't'iyut'e'taan'u' </s>
<s> innale mukhar_jiye sit'iem'aar_aiskaanim'guka_l'_kku' vidhe'yanaakki </s>
<s> iime'khalayil_ var_s'ham'to'r'um' e'zushatamaanam' val'ar_chchayaan'ul'l'atu' </s>
```

8. Automatic Speech Recognition

Automatic Speech Recognition is the process of converting speech into text. Speech recognition systems perform two fundamental operations: Signal modeling and pattern matching [7]. Signal modeling represents process of converting speech signal into a set of

parameters. Pattern matching is the task of finding parameter sets from memory which closely matches the parameter set obtained from the input speech signal [8]. Hence the two important methodologies used in this works are MFCC Cepstral Coefficients [9] for signal modeling and Hidden Markov Model [10] for pattern matching. Mathematically stating computing the probability of a word given a pattern model is computed as the product of two components – acoustic model and language model.[11]

$$\hat{W} = \arg max_W P(Y \mid W) P(W)) / P(Y) \qquad (1)$$

The right hand side of equation (1) has two components: i) the probability of the utterance of the word sequence given the acoustic model of the word sequence and ii), and the probability of sequence of words. The first component P*(Y/W*), known as *the* observation likelihood, which is computed by the acoustic model. The Second component *P(Y)* is estimated using the language model..

8.1 Acoustic model

The Carnegie Mellon University Sphinx-4[12] system is a frame-based, HMM-based, speech recognition system capable of handling large vocabularies. The word modeling is performed based on sub word units (phone set), in terms of which all the words in the dictionary are transcribed. Each phonetic unit considered in its immediate context (which we will refer to as tri-phone) is modeled by 3-state left-to-right HMM model [13] . To reduce the parameter estimation problem, data is shared across states of different tri-phones. These groups of HMM states sharing distributions between its member states are called senones [25]. The acoustic modeling component of the system has four important stages [14] . The first stage is to train the context independent model and then training context dependent models. Decision trees are built on the third stage and finally context independent tied models are created. Here, continuous Hidden Markov models are chosen to represent context dependent phones (tri-phones). The phone likelihood is computed using HMM. The likelihood of the word is computed from the combined likelihood of all the phonemes. The acoustic model thus built is a 3 state continuous HMM, with states clustered using decision tree [15] . The acoustic features used for recognition consist of 39-dimensional acoustic vectors derived every 10 ms spanning an analysis window of 20 ms. The feature vector consists of the first 13 cepstral coefficients (including, the frame energy) and two blocks of 13-dimensional coefficients, one composed of the “delta” cepstral features (velocity) and the other composed of “delta-delta” cepstral features (acceleration). Cepstral mean normalization is always applied at the utterance level to remove statistical biases of the mean which might have been introduced by linear channel distortion.

8.2 CREATION OF LANGUAGE MODEL

The language model employed in our recognition experiments is a tri gram-based language model developed using a training text corpus. These language models are smoothed using the Good-Turing discounting procedure [16].

Importance of a language model in a speech recognition system is vital as acoustic model alone cannot handle the problem of word ambiguity. Word ambiguity may occur in several forms such as similar sounding sounds and word boundaries. With similar sounding sounds, words are indistinguishable to the ear, but are different in spelling and meaning. The words "paat'am' (പാഠം) and " paat'ham' (പാഠo)“are such examples. In continuous speech, word boundaries are also challenging. For instance, the word " talasthaanam' (തലസ്ഥാനം) can be misconstrued as "tala sthaanam' (തലസ്ഥാനം). The use of language model resolves these issues by considering phrases and words that are more likely to be uttered. The trigram based language model with back-off is used for recognition in this work. The language model is created using the CMU statistical LM toolkit [17].

Training and testing is done by n –fold validation techniques. Word Error Rate (WER) is the standard evaluation metric used here for speech recognition. It is computed by SCLITE [18], a scoring and evaluating tool from National Institute of Standards and Technology (NIST)

8.3 Results of Different Speech Recognition Tasks

8.3.1 Isolated Digit Recognition Task

Table 14: Result of Digit Recognition Task

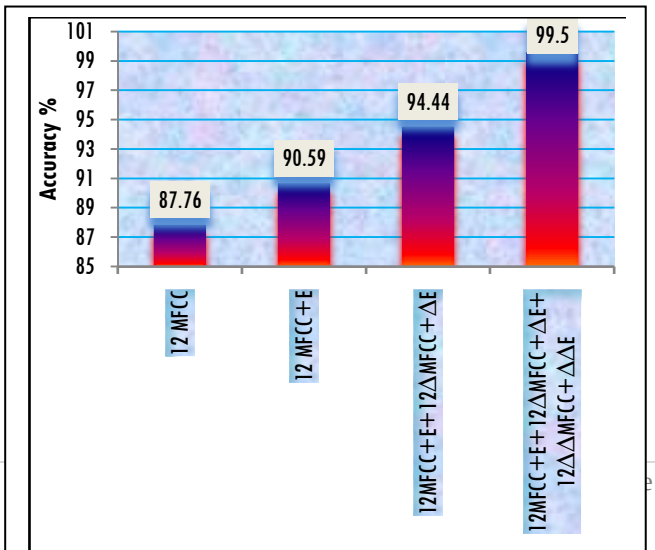
WORD RECOGNITION PERFORMANCE			
Percent Total Error	=	7.1%	(5)
Percent Correct	=	94.3%	(66)
Percent Substitution	=	5.7%	(4)
Percent Deletions	=	0.0%	(0)
Percent Insertions	=	1.4%	(1)
Percent Word Accuracy	=	92.9%	
Ref. words	=		(70)
Hyp. words	=		(71)
Aligned words	=		(71)
CONFUSION PAIRS		Total	(4)
		With >= 1 occurances	(4)
1: 1 -> എട്ട് ==> ഒന്ന്			
2: 1 -> ഏഴ് ==> ആറ്			
3: 1 -> ഏഴ് ==> ഒന്ന്			
4: 1 -> ഒന്പത് ==> ഒന്ന്			

4			

In isolated digit recognition task 92.9% accuracy being obatained as shown in table 14.

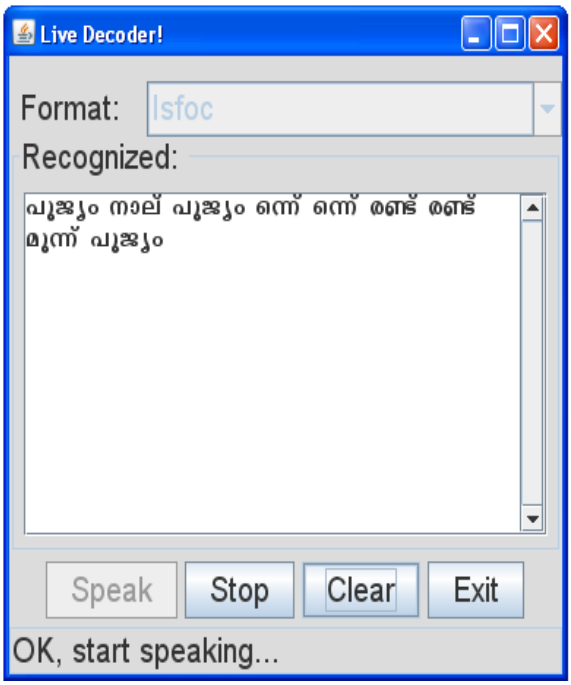
8.3.2 Connected Digit Recognition Tasks

Table 15: Result of Connected Digit Recognition Task



In connected digit recognition task, as per table 15, a very good accuracy is obtained i.e , 99.5 % (with 39 feature vectors) .Table 16 is the snapshot of the live decoder developed for connected digit recogntion task.

Table 16: Live decoder for Connected Digit Recognition Task



8.3.3 Continuous Speech Recognition Task

Table 16 Training and Testing results of Continuous Speech recognition - CDHMM vs. SCHMM Models

Sl.No	Continuous Model		Semi Continuous Model	
	Sentence Recognition Accuracy %			
	Train	Test	Train	Test
1	92.03	86.15	79.2	72
2	91.4	84.11	77.3	70.17
3	90.43	81.13	75	68.18
4	91	85.01	78	70.65
5	90.65	84.36	77.5	70.2
Average	91.102	84.152	77.4	70.24

Speech recognition accuracy of the continuous speech recognition task using continuous and semi continuous model is detailed in table 16. For testing mode, a highest accuracy of 84% is obtained in continuous density hidden Markov model experiment.

9. Summary

Collection of essential resources for the development of speech recognizer for Malayalam language is discussed in this paper . A detailed description on the collection of text corpus and speech corpus for each tasks is being presented. The size of text and speech corpus used for each recognition tasks is detailed. Pronunciation dictionary, Transcription which are created based on these text and speech corpus are explained later with sufficient examples. A complete set of phone list which are prepared in connection with the recognition task is presented in table form. Thus we have created a total speech corpus of size 62000 words and a pronunciation dictionary of size 2480 words. It is anticipated that this work would bridge the start up issues such as collection of speech and text corpus, pronunciation dictionary creation etc, of Malayalam language speech research community to a greater extent.

References :

[1] Sadanandan, S. 1999. Malayalam Phonology: An Optimality Theoretic Approach. Thesis (PhD). University of Southern California.

[2] Mohanan, K.P. and Mohanan, T. 1984. Lexical Phonology of the Consonant System in Malayalam. Linguistic Inquiry, 15, 575-602

[3]Kumari, S.B., 1972. Malayalam Phonetic Reader. Mysore: Central Institute of Indian Languages.

[4]Srikumar, K. and Reddy, N. 1988. An articulatory and acoustic study of trills in Malayalam. Osmania Papers in Linguistics, 14, 42-54.

[5]Radhakrishnan, S. 2009. Perception of Synthetic Vowels by Monolingual and Bilingual Malayalam Speakers . PhD thesis . Kent State University College

[6]Sesma Bailador , Alberto " CorpusCrt Politechnic University of atalonia 1998

[7] L.R.Bahl et.al, A method for the construction of Acoustic Markov Models for Words , IEEE Transactions on Audio, Speech and Language processing, Vol.1,No.4, Oct.1993

[8] B.H. Juang, C.H. Lee and Wu Chou, *Minimum classification error rate methods for speech recognition*, IEEE Trans. Speech & Audio Processing, T-SA, vo.5, No.3, pp.257-265, May 1997.

[9] Rabiner, L. Juang, B. H., Yegnanarayana, B., "Fundamentals of Speech Recognition", Pearson Publishers, 2010.

[10] L.R Rabiner and B.Gold , "Theory and Application of digital Signal processing , Prentice Hall, Englewodd Cliffs , NJ,1975

[11] S. Young (1999). Acoustic Modelling for Large Vocabulary Continuous Speech Recognition. Computational Models of Speech Pattern Processing: Proc NATO Advance Study Institute. K. Ponting, Springer-Verlag: 18-3

[12] http://cmusphinx.sourceforge.net/wiki/tutorial

[13] C.H Lee, L.R Rabinar , R. Pieraccini, and J.G Wilpon , " Accoustic Modelling for Large Vocabulary speech Recognition ", Computer speech and Language , 4:127-165,1990 Vocabulary speech Recognition ", Computer speech and Language , 4:127-165,1990

[14]Levinson, S. E., Rabiner, L. R Sondhi, M. M., (1983). Speaker Independent Isolated Digit Recognition Using Hidden Markov Model, In Proceedings of ICASSP,pp.1049-1052.

[15] S. Young, et. al., the HTKBook, http://htk.eng.

[16] K.F. Lee, Large-vocabulary speaker-independent continuous speech recognition: The Sphinx system, Ph.D. Thesis, Carnegie Mellon University, 1988.

[17]The CMU-Cambridge LM toolkit - http://www.speech.cs.cmu.edu/SLM/toolkit.html

[18] Fiscus, J. (1998) Sclite Scoring Package Version 1.5, US NationalInstitute of Standard Technology(NIST), URL - http://www.itl.nist.gov/iaui/894.01/tools/.