ISSN (Online) : 2454 -7190 **ISSN (Print) 0973-8975**
Copyright reserved © J.Mech.Cont.& Math. Sci., Vol.-14, No.2, March-April (2019) pp 521-540

Modelling South Kamrupi Dialect of Assamese Language using HTK

[1]Ranjan Das, [2]Uzzal Sharma

[1]Research Scholar, Department of Computer Science & Engineering, School of Engineering, Assam Don Bosco University, Azara, Guwahati, Assam, India 781017

[2]Assistant Professor, Department of Computer Science & Engineering, School of Engineering, Assam Don Bosco University, Azara, Guwahati, Assam, India 781017

[1]ranjanmirza@gmail.com, [2]uzzal.sharma@dbuniversity.ac.in

Corresponding Author: Ranjan Das

Email: ranjanmirza@gmail.com

https://doi.org/10.26782/jmcms.2019.04.00038

Abstract

This paper addresses the fundamental issues of developing a speaker independent, dialect modelling system for recognizing the widely spoken, colloquial South Kamrupi dialect of Assamese language. The proposed dialect model is basically designed on Hidden Markov Model (HMM). Hidden Markov Model Toolkit (HTK) is used here as the building block for feature extraction, training, recognition and verification for the model building process. A primary corpus is built as a prerequisite for the empirical study. Altogether, 16 people (9 male, 7 female) are volunteering in the primary corpora building process. The corpora are comprised of one training and two testing sets of recorded speech files. The whole corpora are made up of around 2.5 hours of recordings. The proposed dialect model is trained on South Kamrupi dialect training corpora. A comparative test recognition is carefully designed and carried out which exhibit a recognition correctness of 87.13% for South Kamrupi dialect and 68.52% correctness for the Central Kamrupi dialect. Thus, the findings of this paper evidence that the dialect modelling with proper training has recognized a dialect with better precision.

Keywords : Dialect Modelling, Automatic Speech Recognition, Corpora Building, Feature Extraction, HTK

I. Introduction

Speech is the most fundamental means of communication through which humans express their feelings [XIV]. Speech propagates in the form of acoustic signals. Automatic Speech Recognition (ASR) is a computing discipline, mostly dealing with feature extraction, characterization and recognition of the information present in the

speech signals [XII]. The computing efforts for ASR system can never be successful for a language until the recognition for each and every constituent, colloquial dialectic forms of that language are robustly developed [V]. As the colloquial dialects are those unorganized forms of verbal communication which do not possess any standard written forms, transcribing such a dialect is a complicated situation, however essential for successful machine recognition.

Speech recognition research in Indian languages is still there in their infant states [IV]. Only a few records of computational exploration for the Assamese language are reported so far. Obviously, being a computationally poor language, Assamese does not possess any track record of dialect modelling ever. So, dialect modelling, being an integral part of ASR, is a fundamental necessity but entirely unexplored area of research in Assamese language machine recognition.

The Assamese are the multi-religious, indigenous people, inhabiting over the Bhramaputra valley in the sub-Himalayan range of the state of Assam, India constituting of a diversed mixture of various ethnic races namely, Tibeto-Burmese, Austro-Asiatic and Indo-Aryan, assimilated here, over the thousands of years [I]. The Assamese people can be characteristically identified by the Assamese language and their typical culture. The Assamese language is spoken as the mother tongue by around 16 million Assamese people. In addition to that, around, 35 million use it as lingua franca that are mostly spread in the North-eastern states of India, especially in the states of Nagaland, Arunachal Pradesh and Meghalaya including Assam. Besides, a cognizable number of Assamese speaking people are also available in Bangladesh, Nepal and Bhutan. The constitution of India has recognised the Assamese language as an eighth scheduled, regional language and officially, it is the state language of Assam [XVII].

The Assamese language has an old history which can be traced back to 5th century AD. There are evidences which proclaim that the Assamese language was evolved from Kamrupi dialect in the ancient Kamrupa kingdom [XIX]. In its present form also, Kamrupi is a widely spoken, typical dialectic form of standard Assamese which is rife in the undivided Kamrup area [I]. The undivided Kamrup area is a huge mass of land measuring around 12,384 km^2 inhabited by around 5.24 million people according 2011 census report. Figure 1 shows the geographic map of undivided Kamrup area of the state of Assam, North-east India.

Figure 1. Undivided Kamrup area of the state of Assam, India

As a dialect, the Kamrupi is a heterogeneous, colloquial form and it has a vast range of offshoots spread over the entire undivided Kamrup area. However, it is mostly classified with four major characteristic sub forms namely, Central Kamrupi, West Kamrupi, North Kamrupi and South Kamrupi [XI]. The Central Kamrupi is spoken mainly in the present day Nalbari district centering Nalbari town, the West Kamrupi is spoken in the Barpeta sub-division of the present day Barpeta district centering Barpeta town, the North Kamrupi is spoken in the present day Bajali sub-division of Barpeta district centering Pathsala town and South Kamrupi is spoken in the entire present day Dakhin Kamrup centering Palasbari town. Each of these dialectic forms can be characterised easily by their distinctly identifiable features. In the present study, we are initiating an attempt of dialect modelling for the South Kamrupi dialect of Assamese language only.

Presently, around 1.6 million people speak the South-Kamrupi dialect, who mostly inhabit in the southern bank of the river Brahmaputra of the present day Kamrup district, adjacent to the western part of Guwahati city. Illiteracy is still a major hindrance of this rural folk and this folk belongs to multiple caste, tribe, ethnicity and religion. A huge chunk of these people speak this dialect only and thus, many non-Assamese resident communities of South Kamrup area, such as Marowari, Bihari, Bengali etc. use this dialect as lingua franca. On the other hand, this dialect has enough importance because of its widely prevalent practice. Moreover, it is an integral part of standard Assamese language [XVII].

It is observed that no attempt of developing any computerised dialect modelling system on any of the dialectic forms of Assamese has been reported so far. In order to develop a matured automatic speech recognition system for Assamese language, modelling of each and every dialectic forms of Assamese is mandatory. Computerization in the dialectic level is relatively an unexplored area of research in all major Indian languages [IX]. So, this undertaken study is a pioneering attempt of dialect modelling for Assamese. This study aims to imbibe new impetus to the standard Assamese as a whole. Gradually, this study needs to be extended and integrated to all the other dialectic forms of Assamese for robust machine learning. This research work will benefit every Assamese speaker directly in particular, and every dialect modelling endeavour as a whole.

The rest portion of the paper is organised as follows. Section II describes briefly the Hidden Markov Model(HMM) as the building block for dialect modelling. Here, we are introducing Hidden Markov Model Toolkit(HTK) also as a tool for developing HMM based speech systems. Section III describes the basic framework of the system. Section IV discusses the ASCII transcription used in the modeling for the South Kamrupi dialect. Section V describes the primary corpora building task as an integral part of the empirical experiments and backbone of the dialect model. Section VI is an overall pictorial depiction of the proposed work. Section VII elaborates the steps for preparing the corpora for model building. Section VIII elucidates the model building process followed by training the model. Section IX highlights the design and experiments of the empirical study and a brief discussion on the outcome. Section X concludes the paper with future work.

II. Hidden Markov Model for Dialect Modelling through HTK

A Hidden Markov Model(HMM) is a system of two doubly embedded stochastic processes [XXI], [XIII]. The first process is corresponding to a set of hidden states and the second process is corresponding to a set of visible states. At a certain instant of time, the first process is in some hidden state signifying the time stamp of the system. In the following instance, the hidden state moves to the next hidden state representing the temporal variation. At the same time, a visible state is also generated by the hidden state to signify the spectral variation. Thus, the set of hidden states represents the first stochastic process and the set of visible states represents the other embedded stochastic process. Again, speech can be comprehended as a sequence of acoustic features. There are two types of features. The temporal(time varying) and the spectral(frequency varying). The temporal features are those which have easy physical interpretations such as the energy of signal, zero crossing rate, maximum amplitude, minimum energy etc. On the other hand, spectral features are those which can be obtained by converting the time domain into the frequency domain of the speech signal using the Fourier Transformation such as the fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off etc. These spectral features can be encoded of modeled as a set of fixed sized feature vectors. Some of the well known techniques of speech encoding are Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC), PLP-RASTA (PLP-Relative Spectra) etc.

HMM effectively provides the standards for modelling sequences of temporal spectral vectors [III]. So, all the dialect modelling tasks, being a part of a large vocabulary continuous speech recognition system(LVCSR), can be effectively modelled via HMM. This whole exercise is actually a probabilistic statistical modelling for every phoneme present in the particular dialect. The phonemes are the building blocks of the acoustic model. The principle of HMM is applied here as follows [XII],[XIV]. The captured audio signal is converted into a sequence of fixed size acoustic vector features $X_{1:T} = X_1, X_2, ...X_T$. The Mel Frequency Cepstrum Coefficient (MFCC) feature is used for our encoding task. Now, a decoder finds the sequence of words $\lambda_{1:L} = \lambda_1, \ldots, \lambda_L$ which is likely to have generated by X. So, we have

$$\widehat{\lambda} = arg_{\lambda} max\{P(\lambda|X)\} \quad (1)$$

Applying Bayes' theorem in the above equation, we get

$$\widehat{\lambda} = arg_{\lambda} max\{P(X|\lambda)P(\lambda)\} \quad (2)$$

$P(X|\lambda)$, the conditional probability is determined by the acoustic model whereas $P(\lambda)$ is determined by the language model. A pronunciation dictionary is prepared by describing every word λ by concatenating the phonemes involved into it in the acoustic model. These phonemes are actually extracted during the training phase which incorporates the training speech corpora as well as their respective transcriptions. Thus, the acoustic model is the reference model to identify the unknown utterances. On the other hand, the N-gram model is used for building the

language model. N-gram is simply a sequence of N number of words. So, the N-gram model is a contiguous sequencing of N words(items) from a given sample of speech data. In our language model, the probability of the N-th word is attempted to evaluate by estimating the preceding N-1 words as parameters. It incorporates the transcription of the test corpora.

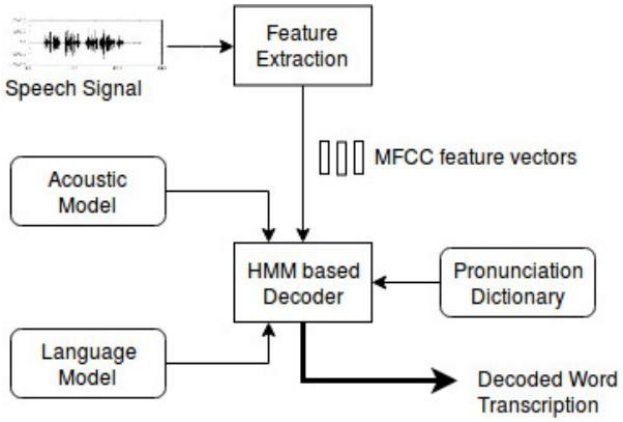

Figure 2. HMM Based Decoding System

Figure 2 describes the basic structure of the HMM-based decoder for word-level transcriptions. The decoding is carried out by matching all possible word sequences for every utterance of the test corpora using the acoustic model. Pruning techniques are used to mitigate the ambiguities. Ambiguities mostly occur in the acoustic model due to variations in pronunciation. This word level transcribing process is continued till we reach the end of the utterances for every recorded speech files of the test corpora. In the end, the word level transcription for the most likely sequences for the whole test corpora is generated which is the output generated by the decoder.

Hidden Markov Model Toolkit (HTK) is used as the backbone for building the HMM-based dialect modelling system in the proposed study [XX]. The HTK is consisting of a set of library modules which provide various utilities for speech analysis, training, testing and results verification. All these library modules are developed using ANSI C programming language and they primarily run on Unix operating system. HTK support both continuous as well as discrete density mixture Gaussians to develop sophisticated HMM based systems [XIII]. HTK based modelling involves two major processing stages, namely : training stage and verification stage. During training stage, training utterances and their associated transcriptions are used to estimate the parameters of a set of HMMs. Baum-Welch algorithm is used in the training phase. On the other hand, Viterbi algorithm decoding technique is used for transcription to recognize the unknown utterance during verification.

The Machine Intelligence Laboratory of Cambridge University Engineering Department(CUED) was the original developer of the HTK. However, during the

nineties of the twentieth century, the copyrights and licenses of HTK were taken over by some commercial enterprises for quite some time. But, at present, CUED has regained back its authority over HTK again. Now, the CUED is taking the responsibility of distributing and upgrading the tool and they are improving it relentlessly. The latest version of the HTK is 3.5 which is used in the proposed modelling task.

III. The Dialect Modelling Framework

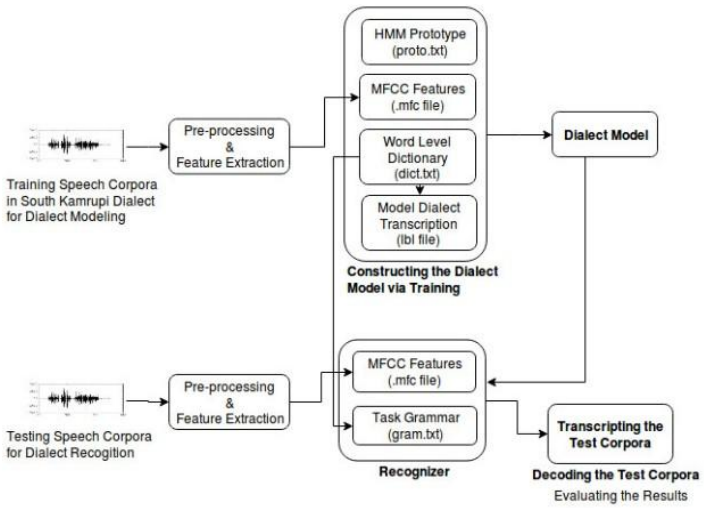

Figure 3. Dialect Modelling System Architecture

Figure 3 describes the dialect modelling system architecture. There are six basic sections subsumed into it [VII]. They are : recording the training and testing corpora as primary data set, pre-processing and feature extraction for both the corpora, HMM-based statistical acoustic dialect modelling by imparting rigorous training, word level as well as phone level transcription based on the training word dictionary, a transcribed output generated by the dialect model for the test corpora and decoding the output for analysis and performance evaluation. However, the basic tasks involved in this system can be broadly classified into two main modules only. They are the training module and the testing module [X]. The training module is responsible for the generation of the system model of the dialect. This system model is a cumulative sum of the South Kamrupi dialect which is developed by incorporating the primary training corpora as the basis. The larger and the more heterogeneous, the training corpora are, the better is the system model. The word level as well as phone level transcription for every unique utterance for the entire training corpora is encoded during this training phase. This system model is the backbone of the recognizer. Contrarily, the testing module is exclusively designed for transcribing the test corpora via the recognizer and then decoding the transcripted output for accuracy and analysis.

IV. ASCII Transcription for South Kamrupi Dialect

The International Phonetic Alphabet(IPA) symbols are linguistically used to describe the transcription for the word pronunciation for every word of any language in their respective dictionary. However, only limited works on the word level pronunciation for the standard Assamese language are reported so far. Works on some different dialectic forms of Assamese language pronunciation are available but the South Kamrupi dialect has no reported evidence as such. Consequently, there is no separate written form for South Kamrupi [XI]. However, in order to carry out the word level transcription for the modelling, every utterances of the training speech corpora must be expressed in the written form. HTK 3.5 supports ASCII based word level transcription labelling only [IX],[XVI]. So, all the transcriptions are carried out manually in the form of text files with .lab extension for every training speech files which have .wav extension individually. We use WaveSurfer as a tool for the purpose. For the ASCII representation, we are using 36 different strings of symbols. It is found that 27 of those symbols are consonants and 9 are vowels. Silences are identified with sil. Different words are separated in a sentence as short pause and it is identified as sp. Observation revealed that অ(a) and আ (aa) are the mostly used vowels; ও(o) is the least used vowel. Again ৰ(r) is the highest used consonant; ঞ(yn/y) is the least used consonant. It is also observed that there are typical pronunciation for the consonant ঙ(ng), which is in many cases pronounced as ৱ (w) also. ঝ(jh) is also typically pronounced in many cases by over-stressing it. The transcription of the South Kamrupi dialect recorded speech files to the ASCII representation is described by the table 1 and table 2.

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
a	aa	i	ii	u	uu	ree	e	oi	o	ou

Table 1. Encoding the Vowels

ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ
k	kh	g	gh	ng/w	s	s	j	jh	yn/y

ট	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ	ন
t	th	d	dh	n	t	th	d	dh	n

প	ফ	ব	ভ	ম	য	ৰ	ল	ৱ	শ
p	ph	b	bh	m	j	r	l	v/w	x

ষ	স	হ	ক্ষ	ড়	ঢ়	য়	ৎ	ং
x	x	h	khy	rr	rh	y	t	ng

Table 2. Encoding the Consonants

V. Building the Corpora

A set of primary corpus are essentially required in the proposed empirical study. The corpora are comprising of the training and the testing corpora separately [VIII],[II]. For building the training corpora, some carefully constructed scripts are prepared first in South Kamrupi dialect as a prototype. 9 different native speakers are volunteering in the training corpora building process out of which 5 are male and 4 are female. All the volunteers are rural, educated and they all fall into the age group ranging from 18 years to 45 years. 25 different sentences and 6 different paragraphs are cautiously assembled to represent the day to day life conversation as well as the typical features of the South Kamrupi dialect. The total sum of 307 words are used in the process. All the sentences are recorded for each volunteer, 3 times each and the paragraphs are recorded for each volunteer, twice each. So, all together, (9 X 25 X 3) 675 numbers of speech files are recorded for the sentences and (9 X 6 X 2) 108 were recorded for the paragraphs for the training corpora building process. There are 5 to 13 different words present in every sentence and around 390 to 860 words are present in every paragraph. Table 3 describes the detail of the training corpora building task.

Number of Volunteers	9 (5 male, 4 female)
Number of sample sentences constructed to represent the South Kamrupi dialect	25 (5 to 13 words)
Number of sample paragraphs constructed to represent the South Kamrupi dialect	6 (390 - 860 words approx.)
Total number of files recorded for the sentences (each sentences thrice each)	9X25X3=675
Total number of files recorded for the paragraphs (each paragraphs twice each)	9X6X2=108
Recording duration(for the sentences) Recording duration(for the paragraphs)	1.859 sec to 5.172 sec 4-7 minutes each(approx.)
Number of Words in the Pronunciation Dictionary	307
Recording sampling rate	44100 Hz
Number of Channels	mono
Recording environment	Noise Free room environment
Recording Hardware	Unidirectional microphone
Recording Software	Audacity

Table 3. Featuring the training Corpora of the South Kamrupi dialect

Again, two different sets of speeches are recorded for the testing corpora building process. This time also, scripts are carefully prepared beforehand. The first test corpora are volunteered by 3 different native speakers of South Kamrupi dialect; 2 male and the other female. The second test corpora are volunteered by 4 different native speakers of Central Kamrupi dialect; 2 male and 2 female. This time also, all the volunteers are rural, educated and they all fall into the age group from 18 years to 45 years. 12 sentences and 2 paragraphs are carefully constructed to represent the day to day life conversation as well as basic characteristics of both the dialects. Each of the sentences and paragraphs is recorded for every speaker; 3 times each. So, in the test corpora of South Kamrupi, there are (12X3X3) 108 recorded files for the

sentences and (2X3X3) 18 recorded files for the paragraphs. Similarly, for the Central Kamrupi corpora, there are (12X4X3) 144 recorded files for the sentences and (2X4X3) 24 recorded files for the paragraphs. Table 4 describes the basic features of the test corpora for South Kamrupi dialect and table 5 describes the basic features of the test corpora for Central Kamrupi dialect.

Number of Volunteers	3 (2 male, 1 female)
Number of sample sentences recorded to represent the South Kamrupi dialect for testing (Each sentences thrice each)	12X3X3=108
Number of sample paragraphs recorded to represent the South Kamrupi dialect for testing (Each paragraphs twice each)	2X3X3=18
File Recording Duration(sentences)	1.897 sec to 4.628 sec
File Recording Duration(paragraphs)	4 to 7 minutes(approx.)
Recording sampling rate	44100 Hz
Number of Channels	mono
Recording environment	Noise Free room environment
Recording Hardware	Unidirectional microphone
Recording Software	Audacity

Table 4. Featuring the test corpora for South Kamrupi dialect

Number of Volunteers	4 (2 male, 2 female)
Number of sample sentences recorded to represent the South Kamrupi dialect for testing (Each sentences thrice each)	12X4X3=144
Number of sample paragraphs recorded to represent the South Kamrupi dialect for testing (Each paragraphs twice each)	2X4X3=24
File Recording Duration(sentences)	1.823 sec to 4.783 sec
File Recording Duration(paragraphs)	4 to 7 minutes(approx.)
Recording sampling rate	44100 Hz
Number of Channels	mono
Recording environment	Noise Free room environment
Recording Hardware	Unidirectional microphone
Recording Software	Audacity

Table 5. Featuring the test corpora for Central Kamrupi dialect

All the speech files are captured using uni-directional microphone SONY ICD-UX533F with the recording tool Audacity in the .wav file format [VI]. A noise free room environment is enabled during the entire speech capturing process. An approximate distance of 5-10 cm is maintained between the mouth of the volunteers and the microphone. The sampling rate of all the recorded .wav files are assigned to be 44100 Hz and an accuracy of 16 bits/sample is maintained in the PCM wave format. The number of channels for all recordings is mono. The time duration of all the recorded files is in between 1.823 sec to 5.172 sec for the sentences and 4 to 7 minutes approximately for the paragraphs. Almost 2 and a half hours of recorded

speech files are constituting the whole corpora which encompass both, the training as well as testing corpora collectively.

VI. Proposed Work

The entire corpora comprises of 3 corpus as mentioned in the previous section. All the individual recorded speech files of the prepared corpora are employed in the proposed work systematically. The tasks involve in the proposed dialect modelling system are organised into 3 basic phases, namely : Phase 1 : Preparing the data, Phase 2 : Model building and training, and Phase 3 : Evaluation. Figure 4 depicts each of these phases along with the subtasks undertaken within each one of them.

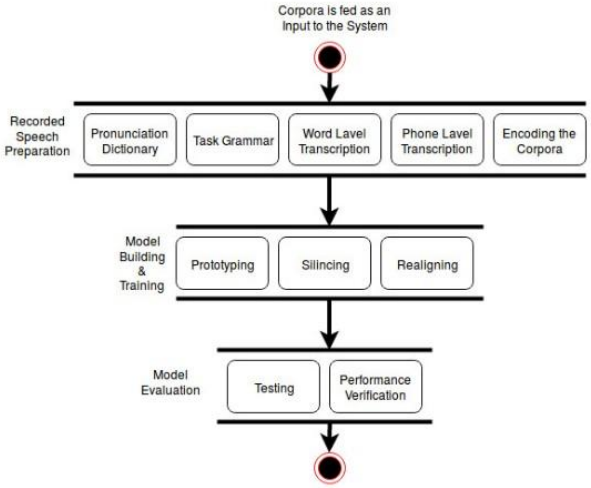

Figure 4. The Activities undertaken in the Dialect Modelling

HTK is used as a building block of the proposed dialect modelling task. At each of the phases, various tools of HTK are used by preparing shell script files independently which are executed in the traditional command-line style using Unix terminals. All these tools require a number of arguments as well as optional arguments. The optional arguments are always prefixed by a minus sign. The arguments are various types of text files with varied extensions. Few of these text files are manually prepared ; and the others are intermediate files prepared by the shell script execution. The following section describes each of these phases; one by one.

VII. Preparing the Recorded Speech Files

There are some mandatory preparations which are to be accomplished on each and every file of the recorded corpora [XI]. Various tools of HTK are successively used to create multiple intermediate files of different formats which ultimately, in turn, will participate in the modelling of the dialect. The data preparation in our application basically comprises of the creation of the following files.

Pronunciation Dictionary

As the entire speech recording process is navigated by some carefully prepared scripts, each and every word of the recorded speech files are enlisted from those scripts one by one manually using the phoneme encoding scheme of South Kamrupi (Assamese) to ASCII in the form of a text file for the training corpora. This file is termed as pronunciation dictionary. The UNIX commands uniq and sort are used to sort the entries of the pronunciation dictionary uniquely. Every word is expanded as a sequence of multiple phones here which are separated by short-pause(sp). Silence is also incorporated into this dictionary as sil for the beginning and end of each of the training speech files. There is the possibility of multiple pronunciations for some of the words. So, multiple entries are entered for these variably pronouncing words with the varied sequence of phonemes in our dictionary.

The Task Grammar

Task grammar is a logical representation of the entire set of words present in the pronunciation dictionary for the user to visualise the whole training corpora. There is a grammar definition language in HTK which consists of a set of variable definitions followed by a regular expression describing every word present in the training corpora. So, the complete task grammar represents a high level, the abstract network of words for the model. At the same time, an explicit word network is actually necessary for the HTK to model. This explicit, physical network encompasses word-to-word transition from every word to every other word present in the training corpora. This physical network is called wordnet. This wordnet is built by using the HTK tool HParse from the task grammar.

Word Level Transcription

There are multiple recorded files (.wav extension) in the training as well as testing corpora. All these speech files are transcribed to their equivalent set of encoded words using the list of South Kamrupi (Assamese) to ASCII representation one by one [IX]. The open source software WaveSurfer is used manually for this purpose. WaveSurfer is a customizable audio waveform visualization editor and it is widely used for interactive displaying of the sound pressure waveforms, spectral sections, spectrograms, pitch tracks, transcription etc. Each recorded speech files are represented in a respective text label file with the extension .lbl and with the same names as the recorded speech files. Every word is separated by newlines in these text files. Again, all these text files are also assembled together separately in a collective text file, identified as Master Label File (.mlf). Different master label files are created for training and testing corpora separately.

Phone Level Transcription

Form the word level transcription of the training corpora, the phone level transcription file for the training corpora is created. This phone level transcription file is also a master label file with the same extension name .mlf likewise. The HTK tool HLED is used for this purpose. An edited script file with the extension .led is employed here in this exercise. This script file mostly contains commands to replace the words from word level transcription to phone level transcription. The

pronunciation dictionary is also engaged in the operation. This newly created phone level transcription is acting as the operational basis for training the proposed dialect modelling task.

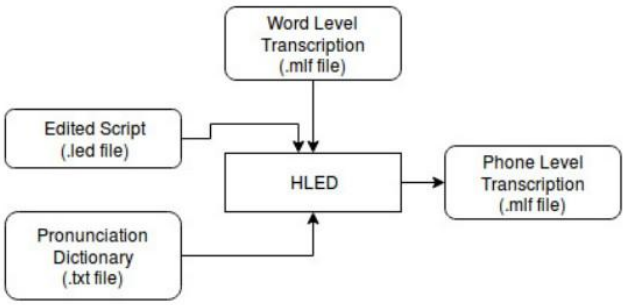

Figure 5. Generation of Phone Level Transcription

Encoding the Corpora

The main purpose of encoding the recorded speech files is the acoustic feature extraction [IX]. Acoustic features are those numerical parameters of an audio signal which characterize the audio segment with its static aspects in a short succession of time instance. Feature extraction is the process of computing the acoustical features. During feature extraction, only the relevant pieces of information for the intended study are retained. The superfluous information is often discarded. Mel Frequency Cepstral Coefficient (MFCC) is the most relevant acoustic feature vector which can be extracted for the dialect recognition purpose using HTK. The total number of MFCC features extracted here is 39 out of which 13 are static vectors, 13 are delta coefficients and 13 are acceleration coefficients. HTK enhances the performance of the model by adding time derivatives to the static vectors [XI]. The first order regression coefficients with respect to time are termed as delta coefficients and the second order regression coefficients with respect to time are termed as acceleration coefficients or delta-delta coefficients.

Target file type	MFCC file(.mfc extension)
Source file type	Recorded speech files (.wav extension)
Assumed Energy component	C_0
Frame period	10 msec
Number of Filter Banks	26 Channels
The window used	Hamming window
Coefficient for computing first order pre-emphesis	.97
Energy normalization	yes
Live audio system	no
Output compression	yes
Number of MFCC coefficients as output	39(static vector 13, delta coefficients 13, acceleration coefficients 13)

Table 6. Encoding the Speech Configuration

The HTK tool HCOPY is used for coding all the recorded speech files enlisted in the training as well as testing corpora [XX]. This tool HCOPY generates MFCC feature files with the extension name .mfc from every speech files individually. A configuration file with the extension .config is specially created for maneuvering the encoding process. The table depicts the basic features of the configuration file which is responsible for encoding every recorded speech files of the whole speech corpora.

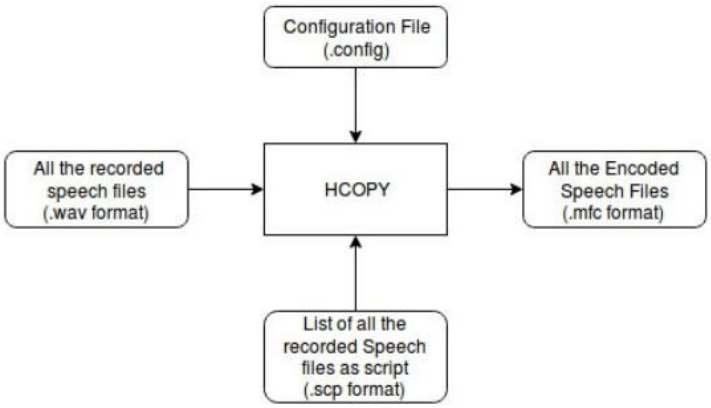

Figure 6. Encoding the recorded files to mfcc

VIII. Model Building and Training

This section primarily discusses the development of the proposed dialect modelling system for the South Kamrupi dialect which is standing on a single-Gaussian HMM. The process is beginning with a prototype building which comprises of a set of indistinguishable HMMs. The mean and the variance for the whole set of the mono-phones are the same at this stage. This prototype is repeatedly trained and gradually silence model and short-pause model are also incorporated into it as extensions. The process of retraining and rebuilding is continued in order to address the issue of multiple pronunciations for some of the entries of the pronunciation dictionary.

Prototyping

A prototype is the preliminary version of an entity from which other sophisticated structures can be developed by successive iteration. In the proposed dialect modelling system, a prototype model is defined first for the HMM training as the model topology [XX]. It is a 3 state HMM where each ellipse vector is of length 39. All the mean values are assumed to be 0.0 and the variance values are assumed to be 1.0 for the prototype initially.

HCOMPV is an HTK tool which is used to compute the global mean and variance for all the HMMs and a new version of the prototype is derived from the assumed prototype. A list of all the .wav files for the training corpora is passed as a script file

to the HCOMPV tool. A Master Macro File(MMF) called hmmdefs is generated in the process which contains all the HMMs along with the word level labellings including the silence model. However, this file does not possess the short-pause model for the entire training data. A macros file, which is a text file is also created in the computation which contains the variance vector for the first version of HMM.

These hmmdefs and macros are passed through a different HTK tool called HREST as the first version of mono-phone HMMs to compute the next level of HMMs(hmmdefs and macros). The phone level transcription file is also incorporated in the process. The forward-backward algorithm with a pruning threshold value of 400.0 is used for the processing to reduce the computational cost. HEREST has applied 3 times again in the application in a recursive way which re-estimates the model HMM(hmmdefs and macros).

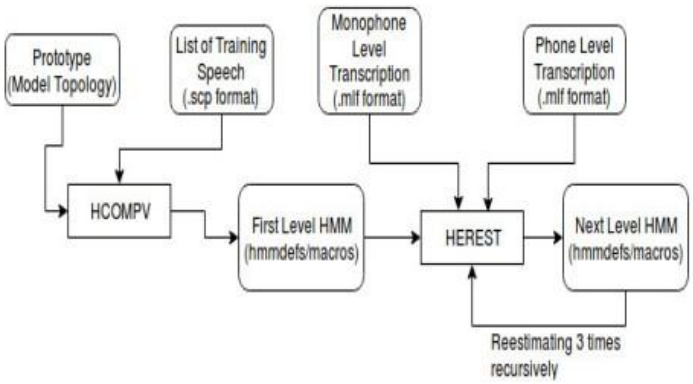

Figure 7. HMM generation via re-estimation

Silencing

The model is now refined by allowing the absorption of various impulsive noises and it is integrated with the short-pause(sp) also. An HTK tool HHED is used for the purpose. Now, the tool HEREST is applied recursively twice again on the model. Thus, mono-phone HMM model building process is re-estimated with more clarity.

Realigning

As there are multiple pronunciations for some of the words present in the training pronunciation dictionary, we are realigning the model again for better phone level transcription. The HTK tool HVITE is used for realignment which makes our transcription more robust. Now, all pronunciations for each word are incorporated in our model.

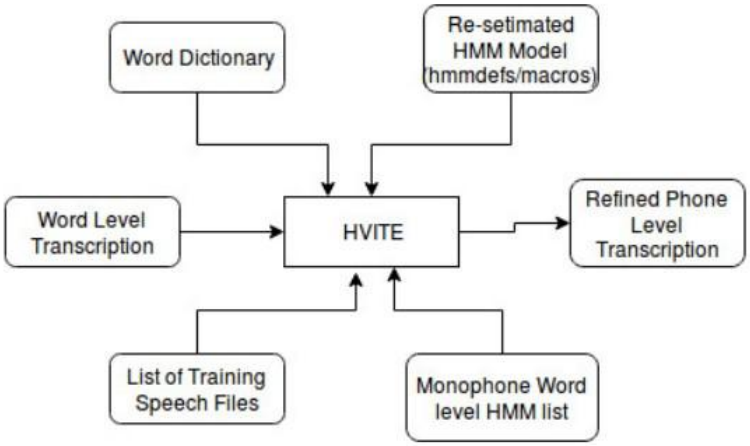

Figure 8. Transcription Refinement from Word level to Phone level

HEREST has again applied twice recursively to re-estimate the parameters of the model with the refined phone level transcription file. This new phone level transcription is generated in the previous step by using the tool HVITE. Thus, the training and building exercise of the model is completed after 9 rounds of successive computation and recursive re-estimation.

IX. Evaluating the Model

Now, the computing dialect model is ready for evaluation as a recognizer. There are two sets of already prepared test corpora; one for the South Kamrupi dialect and the other for the Central Kamrupi dialect. Both these corpora are encoded separately using the HTK tool HCOPY. The list of all the encoded files for testing of the South Kamrupi dialect are enlisted in a script file testSouthKamrupi.scp and for the testing of the Central Kamrupi dialect are enlisted in another script file testCentralKamrupi.scp.

Testing

The HTK tool HVite is used for the dialect recognition task also. HVite is applied separately on testSouthKamrupi.scp and testCentralKamrupi.scp successively which generates two separate word-level transcription output files recoutSouthKamrupi.mlf and recoutCentralKamrupi.mlf respectively. Both these output files are Master Label File. There are issues arising on scalability which are taken care of by the tool Hvite itself. The generation of these word-level transcription output files for the South Kamrupi dialect and the Central Kamrupi dialect is the main job of the proposed model. Thus, the proposed automatic speech recognition task is completed at this juncture. However, the enumeration of the performance is imperatively necessary in order to validate the generated output of the model.

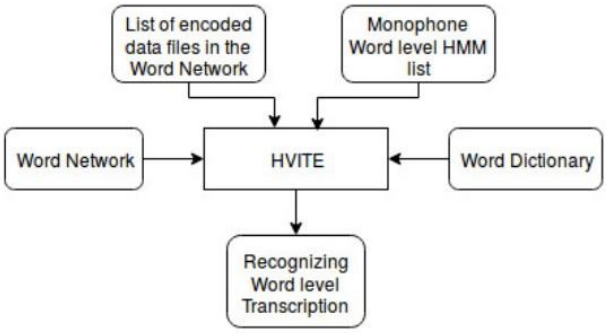

Figure 9. Recognizing the test corpora by generating word-level transcription

Performance Verification

The performance represents the act of accomplishment of a task or thing. Verifying the performance is intended for building a system to access the performance. Figure 10 describes the building block for designing the performance verification for our model.

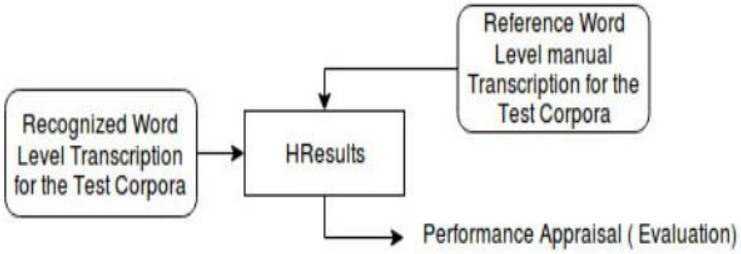

Figure 10. Evaluating the Output

During the training phase of the word level transcription, manual labeling task was carried out for both the test corpora as well, along with the training corpora using the tool WaveSurfer. All those labeled files (.lbl extensions) for both the test corpora are stored in two separate Master Label Files, namely testrefSouthKamrupi.mlf for the South Kamrupi dialect and testrefCentralKamrupi.mlf for the Central Kamrupi dialect respectively. Now, for the practical performance appraisal, the HTK tool HResults is applied to both these files separately. Firstly, for the South Kamrupi dialect which involves the files recoutSouthKamrupi.mlf and testrefSouthKamrupi.mlf. Secondly, for the Central Kamrupi dialect which involves the files recoutCentralKamrupi.mlf and testrefCentralKamrupi.mlf.

The table 7 and table 8 describes the comparative performance of our model.

Output file of word-level transcription generated by the model for the South Kamrupi test corpora	recoutSouthKamrupi.mlf
Manually prepared reference file of word level transcription	TestrefSouthKamrupi.mlf
Number of utterances in the South Kamrupi test corpora (N)	262
No. of correctly recognized Utterances in the Test Corpora(H)	227
Percentage of accuracy (ACC)	86.64%
Total number of Words in the South Kamrupi test corpora (N)	6232
Total number of correctly recognized Words in the South Kamrupi test corpora (H)	5431
Correctness in percentage(COR)	87.13%

Table 7. Recognition details of South Kamrupi dialect

Output file of word-level transcription generated by the model for the Central Kamrupi test corpora	recoutCentralKamrupi.mlf
Manually prepared reference file of word level transcription	TestrefCentralKamrupi.mlf
Number of utterances in the Central Kamrupi test corpora (N)	272
No. of correctly recognized Utterances in the Test Corpora(H)	184
Percentage of accuracy (ACC)	67.17%
Total number of Words in the Central Kamrupi test corpora (N)	8349
Total number of correctly recognized Words in the Central Kamrupi test corpora (H)	5721
Correctness in percentage(COR)	68.52%

Table 8. Recognition Details for Central Kamrupi dialect

Being the dialects of the same language, both, the South Kamrupi dialect and the Central Kamrupi dialect possess morphologically identical structures. The variation in utterance level is the only point of disparity between them [XV],[XXII]. The results obtained shows that the correctness of the recognition for South Kamrupi dialect is 87.13% whereas, for Central Kamrupi, it is just 68.52%. Observation reveals that the extent of accuracy is lower than the correctness which is 86.64% in case of South Kamrupi dialect and 67.17% case of Central Kamrupi dialect. This is because the insertion error is ignored in either case for correctness. It is also observed that when the recorded speech files for the paragraphs (108 number of files) are dropped from the training corpora in the same test-bed of the proposed dialect modelling, the correctness for South Kamrupi recognition drops dramatically to 83.27% and for Central Kamrupi recognition, it rises to 71.43%. Thus, this observation infers that a robust, rigorous training is prerequisite for better precision of the results.

X. Conclusion and Future Work

This is a successful automatic speech recognition endeavour for developing a speaker independent, abstract, off-line model of a spoken dialect of Assamese language. No attempt of dialect modelling has ever been reported in any of the colloquial dialectic forms of Assamese language so far. In that sense, it is a pathfinding effort. The development of the primary corpora, their transcriptions and the dialect model building are the most significant parts of this research work. The empirical detail of the works has been described in this paper which exhibits satisfactory outcomes. The precision of the model building task are expected to improve by enlarging the corpora; both for the training as well as testing. This research can further be extended by incorporating tied-state triphone in the transcribing process. The model can be enhanced further in more pragmatic and lively way by inserting real-time spoken speech data stream online as input in the corpora. Of course, an interactive interface has to be developed in that case for the ease of the user. Finally, this work is the foundation which will certainly benefit the people of South Kamrup directly in the days to come. As a whole, this computing task is an epoch-making contribution to the Assamese language research, particularly in the field of dialect modelling, which has sufficient scope to upgrade as realistic applications in the near future.

References

I. B. Kakati, “Assamese its formation and development”. Guwahati, India, LBS publication, 2007.

II. B. Ramani, S. L Christina, G. A Rachel, V. S Solomi, M. K Nandwana, A. Prakash, S. A Shanmugam, R. Krishnan, S. K Prahalad and K.Samudravijaya, “A common attribute based unified hts framework for speech synthesis in Indian languages”, In Eighth ISCA Workshop on Speech Synthesis, 2013.

III. D. Jurafsky and J. H Martin, “Speech and language processing”, volume 3. Pearson London, 2014.

IV. D. S Kulkarni, R. R Deshmukh, P. P Shrishrimal, and S. D Waghmare, “Htk based speech recognition systems for indian regional languages: A review” 2016.

V. G. Aneeja and B. Yegnanarayana, “Extraction of fundamental frequency from degraded speech using temporal envelopes at high snr frequencies”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4):829–838, 2017.

VI. G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P Singh, RNV Sitaram, and SP Kishore, "Development of indian language speech databases for large vocabulary speech recognition systems", In Proc. SPECOM, 2005.

VII. G. Salvi, "Htk tutorial", KTH Royal Institute of Technology, Department of Speech, Music and Hearing, Drottning Kristinas, 31, 2003.

VIII. H. Sarfraz, S. Hussain, R. Bokhari, A. A Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed and R. Parveen, "Speech corpus development for a speaker independent spontaneous urdu speech recognition system", Proceedings of the O-COCOSDA,Kathmandu, Nepal, 2010.

IX. H. Sarma, N. Saharia, and U. Sharma, "Development and analysis of speech recognition systems for assamese language using htk", ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(1):7, 2017.

X. K. Kumar, RK Aggarwal, and A. Jain, "A hindi speech recognition system for connected words using htk", International Journal of Computational Systems Engineering, 1(1):25–32, 2012.

XI. K. Medhi, "Assamese grammar and origin of the Assamese language". Publication Board, Assam, 1988.

XII. K. Tokuda and H. Zen, "Fundamentals and recent advances in hmm-based speech synthesis", Tutorial of INTERSPEECH, 2009.

XIII. L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under- resourced languages: A survey", Speech Communication, 56:85–100, 2014.

XIV. L. R Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", Proceedings of the IEEE, 77(2):257–286, 1989.

XV. M. Dua, RK Aggarwal, V. Kadyan and S. Dua, "Punjabi automatic speech recognition using htk", International Journal of Computer Science Issues (IJCSI), 9(4):359, 2012.

XVI. M. S Liang, R. Y Lyu, and Y. C Chiang, "Phonetic transcription using speech recognition technique considering variations in pronunciation", In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 4, pages IV–109. IEEE, 2007.

XVII. R. Das and U. Sharma, "Extracting acoustic feature vectors of south kamrupi dialect through mfcc", In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pages 2808–2811. IEEE, 2016.

XVIII. S. L Maguer, I. Steiner, and A. Hewer, "An hmm/dnn comparison for synchronized text-to- speech and tongue motion synthesis", Proc. Interspeech 2017, pages 239–243, 2017.

XIX. S. Mahanta. "Assamese", Journal of the International Phonetic Association, 42(2):217–224, 2012.

XX. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, and D. Povey, "The htk book", Cambridge university engineering department, 3.5:433, 2015.

XXI. T. F Quatieri, "Discrete-time speech signal processing: principles and practice", Pearson Education India, 2006.

XXII. V. Sneha, G Hardhika, K J. Priya, and D. Gupta, "Isolated kannada speech recognition using htk —a detailed approach", In Progress in Advanced Computing and Intelligent Engineering, pages 185–194. Springer, 2018.