Project Title: "To develop an Automatic Speech Recognition (ASR) system for Manipuri dialects."

Name & Designation of Principal Investigator & Co-Investigator:

Principal Investigator

Prof. Khumanthem Manglem Singh

Professor Department of Computer Science and Engineering

NIT Manipur, Langol, Imphal West, Manipur-795004

Co-Investigator

Dr. Yambem Jina Chanu

Assistant Professor Department of Computer Science and Engineering

NIT Manipur, Langol, Imphal West, Manipur-795004

DST Sanction Order No. and date: *6/2/2017-S&T dtd 29/11/2017 31st March, 2018*

Project Serial Number (as per DST Sanction Order) :

Sanctioned Project cost and duration: Rs 1 Lakh, 12 months

Actual Project cost and duration: Rs 1 Lakh, 12 months

Date of Project start and completion: 1st November, 2018 to 31st March, 2019

ABSTRACT

Speech is the most common means of communication to interact with one another. Speech is the spoken version of natural language. Automatic speech recognition (ASR), translating of spoken language into readable text in real time, is still a challenging task due to high viability in speech signals. Deep Neural Network is becoming a mainstream technology for Automatic speech recognition. The main target of this project it to develop an ASR system for Manipuri dialect namely for Kakching and Andro dialect. Data for Kakching and Andro dialect is collected from native speaker. Finally, results are tested and compare between Kakching and Andro dialects.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

:

FIRST PAGE :

1. Project Title: "To develop an Automatic Speech Recognition (ASR) system for Manipuri dialects."
2. Name & Designation of Principal Investigator & Co-Investigator:

Principal Investigator

Prof. Khumanthem Manglem Singh

Professor Department of Computer Science and Engineering

NIT Manipur, Langol, Imphal West, Manipur-795004

Co-Investigator

Dr. Yambem Jina Chanu

Assistant Professor Department of Computer Science and Engineering

NIT Manipur, Langol, Imphal West, Manipur-795004

3. DST Sanction Order No. and date: *6/2/2017-S&T dtd 29/11/2017 31st March, 2018*
4. Project Serial Number (as per DST Sanction Order) :
5. Sanctioned Project cost and duration: Rs 1 Lakh, 12 months
6. Actual Project cost and duration: Rs 1 Lakh, 12 months
7. Date of Project start and completion: 1st November, 2018 to 31st March, 2019

Signature of the Investigators: **Signature of the Co Investigators:**

(Prof. Khumanthem Manglem Singh**)** **(**Dr. Yambem Jina Chanu**)**

1. Project proposal details as per formats PART 1 to PART5

R&D Project proposal : PART1: IDENTIFICATION

1.	Project Title :	"To develop an Automatic Speech Recognition (ASR) system for Manipuri dialects"
2.	Scheme applied for :	DST-Manipur Short-term R&D project
3.	DST Thrust Area for Short-term R&D projects (as listed in guidelines) :	Projects with the aim of addressing requirements of exclusive concern to Manipur using S&T inputs.
4.	Project duration : (not more than one year)	1 Year
5.	Total project cost :	Rs. 1 Lakh
6.	Principal investigator **:**	
	Name :	Prof. Khumanthem Manglem Singh
	Designation:	Professor
	Organisation:	National Institute of Technology Manipur
	Address for correspondence:	Langol, Imphal – 795 004, Manipur
	Mobile No.	08837487361
	Email:	manglem@gmail.com
7.	Co-Investigator :	
	Name :	Dr. Yambem Jina Chanu
	Designation:	Assistant Professor
	Organisation:	National Institute of Technology Manipur
	Address for correspondence:	Langol, Imphal – 795 004, Manipur
	Mobile No.	08974465178
	Email:	jina.yambem@gmail.com
8.	Self attested Passport size Photograph of Principal Investigator:	Self attested Passport size Photograph of Co-Investigator:
	Signature:	Signature:

R&D Project proposal : PART 2 : SUMMARY OF PROJECT

1	Name of Institution:	National Institute of Technology Manipur
2	Principal investigator & Co-Investigator	Prof. Khumanthem Manglem Singh (PI) & Dr. Yambem Jina Chanu (Co PI)
3	Project Title:	"To develop an Automatic Speech Recognition (ASR) system for Manipuri dialects"
4	Project Objective :	The objective is to develop an Automatic Speech Recognition(ASR) system for Manipuri dialects.
5	Outcome of the Project :	Once this project is implemented it could be further extended to other Manipuri dialects. At the end it will help in building an ASR system for Manipuri language which can recognize various dialects exist in Manipur. Due to time constraints at first the focus will be given to Kakching and Andro dialects.
6	Relevance of the outcome to socio-environment / economic development of the people of Manipur.	The project could become the base model for developing many ASR system like Hospital ASR where the patients from different part of Manipur could call the system and help them in finding the schedule of a particular Hospital like whether the Doctor they prefer is on duty today or not vice versa, as Manipur road connectivity is very poor compared to other states in India, which is being encourage by the naturally constructed geography of Manipur and it is constructed in such a manner that valley region is surrounded by 9 range of hills. On top of this if the connectivity is not good then it really affects the life of every common people in various ends. Hence if Hospital ASR is developed then it can reduce unnecessary visits by a patient to the particular hospital if their doctor is not available on the said day.
7	Work plan / Methodology :	
8	Proposed budget and project duration	Rs. Lakh and 1 year
9	Any special point of significance.	This project will become the base model for various ASR system of various real time application in Manipuri language which includes varieties of existing dialects of Manipur.

Duration	Description
1-2 months	➢ Manpower recruitment and training. ➢ Data collection.
3-6 months	➢ Segmentation and Features Extraction.
7-10 months	➢ Development of Classification model
11-12 months	➢ Post processing and trials

Signature of Principal Investigator

(Prof. Khumanthem Manglem Singh**)**

R&D Project proposal : PART – 3 : TECHNICAL DETAILS

Project Title:

"To develop an Automatic Speech Recognition (ASR) system for Manipuri dialects"

1. Introduction:

1.1. Origin of the Proposal :

A Speech Recognition project like ASR for agricultural commodities price accessing system[1] is being developed in IIT Guwahati. Knowing the existence of such system it would be helpful if we could have the same for Manipuri language for our state. But as we know many researchers have started some works related ASR for Manipuri language but no researchers have taken care of the existing dialects of Manipur which could become a part of set back for the research community if we didn't start it now. Speaking of which Manipuri language which is spoken by 1,648,000 [2] is a tonal language with different dialects like Phayeng, Sekmai, Andro, Kakching, Koutruk, Kwatha etc. [3,4]. To better understand and tone recognition of the language, a deep research is required. Due to this reason we are motivated to choose such field as a project. Currently due to time constraints only two dialects (Kakching and Andro) will be selected.

1.2 Definition of the project:

The project will be implemented in such a manner that the system will recognize the two dialects of Manipur namely, Kakching and Andro.

1.3 Objectives of the project:

The main objective is to develop an Automatic Speech Recognition System(ASR) for Manipuri dialects (Kakching and Andro).

1.4 Science Technology content of the proposal :

A typical speech recognition system is developed with major components that include acoustic front-end, acoustic model, lexicon, language model and decoder as shown in Figure 1. Acoustic front-end takes care of converting the speech signal into appropriate features which provides useful information for recognition. The input audio waveform from a microphone is converted into a sequence of fixed-size acoustic vectors is a process called feature extraction. The parameters of word / phone models are estimated from the acoustic vectors of training data. The decoder operates by searching through all possible word sequences to find the sequence of words that are most likely to generate. The likelihood is defined as an acoustic model P(O/W) and P(W) is determined by a language model [Annexure A].

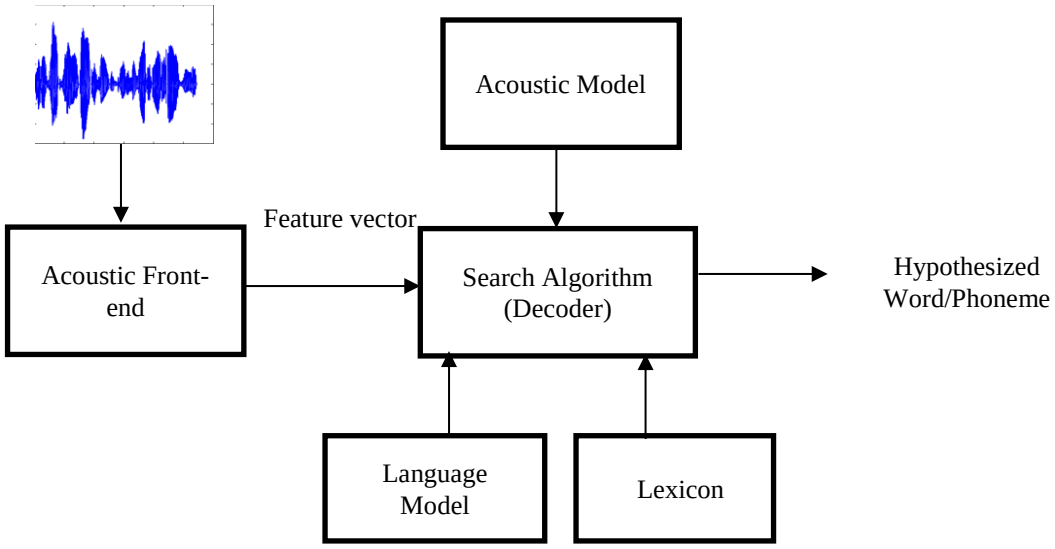


Figure 1: Speech Recognition Architecture

The functionality of automatic speech recognition system can be described as an extraction of a number of speech parameters from the acoustic speech signal for each word or sub-word unit. The speech parameters describe the word or sub-word by their variation over time and together they build up a pattern that characterizes the word or sub-word. In a training phase the operator will read all the words of the vocabulary of the current application. The word patterns are stored and later when a word is to be recognized its pattern is compared to the stored patterns and the word that gives the best match is selected. This technique is generally referred to as pattern recognition.

1.5 Importance of the proposal with reference to Manipur.

With reference to Manipur this project will act as the base model for various ASR system of Manipuri language including different dialects of Manipur.

2. Review status of the subject:

2.1 International status:

In International platform no such system has been developed.

2.2 National status:

According to Ministry of Electronics and Information Technology consortium project on ASR system for various language like Assamese from NE region is already implemented but not for Manipuri language[5].

2.3 Importance of the project in the context of current status:

Need/rationale for taking up the proposed project in the context of current status As per the information available till date no such project has be taken up so far which leads to urgent requirement for such system specially giving preference to dialects (Kakching and Andro due to time constraint of the project) existing in Manipur.

3. Capability of the Organization:

3.1 Specialists consulted/ to be consulted: Nil

3.2 Expertise available with the Investigating group:

In the Department of Computer Science and Engineering various faculties with different specializations are jointly serving in NIT Manipur. So faculties are ready for any consultant.

3.3 List of on-going and completed projects of this group with the following details:

Sl. No.	Agencies	Grant/Amount mobilized (Rs in Lacs)	Period
1.	DIT MIT	2.93 Crore	2009-2010
2.	DIT MIT	18	2005-2010

4. Work-plan :

4.1 Methodology and experimental set-up to be adopted:

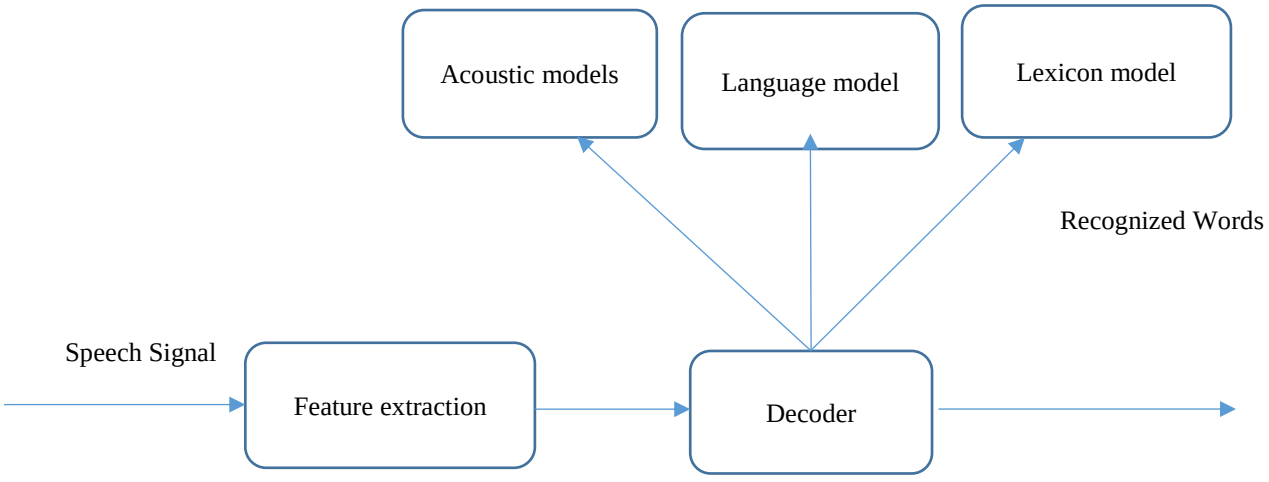


Figure 2: Block diagram of ASR system

The block diagram of the proposed system is shown in Figure 2.

4.2 Materials and data to be collected and examined: All data are collected for Kakching and Andro native speakers.

4.3 Method of analysis and conclusion:

This ASR systems use Deep Neural Networks (DNN) for good accuracy and robustness [48]. This research used scripts provided with Kaldi toolkit [49] for training DNN-based ASR systems, and IRSTI.M tool [50] for building language models. Kaldi is based upon finite state transducers and it is compiled against the OpenFst toolkit [51]. Results of the proposed model will be analyse along with the evaluated word error rate (WER) of the system under different scenarios and compare with the result of the baseline system. Also compare the results between the two Manipuri dialects namely Kakching and Andro.

4.4 Time schedule of activities giving milestones:

- Task 1-Data collection.
- Task 2-Segmentation and Features Extraction.
- Task 3-Development of Classification model
- Task 4-Post processing and trials

Figure 3: Work plan of the proposal. Duration is in months.

5. Outcome and Assessment :

5.1 Nature of outcome of the project :

An Automatic Speech Recognition(ASR) system which can recognize Kakching and Andro dialects.

5.2 Anticipated contribution from the project towards increasing the state and knowledge on the subject. :

Once such project is successful then different linguist can participate for different dialect available in the state.

5.3 Proposed academic benefits from the project in terms of number of research publications and manpower trained.:

By publishing the papers scholars can really know how the ASR model for Manipuri dialects revolve in technical front.

5.4 Anticipated practical benefits resulting from the outcome/financings of the project.

In future an ASR for real time application in Manipuri language could be developed.

5.5 Anticipated practical benefits which are relevant particularly to the socio-economic development of the people of Manipur.

Currently viewing the financial condition of Manipur, women are the main source for handloom or weaving industry with small investment. So once this project is successful then it will pave the path for easy communication among the foreign tourist with local handloom promoter in order to promote their product and as the number of foreign visitors visiting Manipur is on high rise as well as the likes of indigenous handloom product is also on high rise which will help the women entrepreneur in weaving industries of small investment economically.

5.6 Names and addresses of experts/ institutions interested in the project outcome of the project: Nil

5.7 Whether Patent is proposed.: NIL

R&D Project proposal : PART – 4 : BUDGET ESTIMATES

1. TOTAL BUDGET:

Item	Budget Estimate(in Rs)
Contingencies	4000/-
Total :	**4000/-**

2. DETAILS OF MANPOWER:

Designation of manpower	Qualification	Monthly honoraria	No. of Months to be engaged
Project Assistant	B.Tech CSE/IT	16,000/-	6
Total :			**96,000/-**

JUSTIFICATION :	Please give justification for budget items.
Monthly honoria of Rs 16,000/-	a) 16,000 * 6= 96000/-

Note: Manpower will normally be treated as part-time engagement and the honoria is to be proposed accordingly.

R&D Project proposal : Part 5 : BIODATA OF INVESTIGATORS

(Separately for Principal Investigator/ Co-Investigator)

A. Name & Designation:	Prof. Khumanthem Manglem Singh Professor
B. Institution	National Institute of Technology Manipur Langol, Imphal
C.: Date of Birth :	21-01-1963
D. Whether belongs to SC/ST/OBC :	General
E. Academic & Professional career : **Academic career :** **Professional career :**	 29 years 29 years
F. Title of Doctoral thesis :	"New Development in Median Filter"
G. Award/Prize/Certificate etc. won by the Investigator :	NIL
H. Publications : Books : Research Papers:	 1 number 130 numbers

Prof. Khumanthem Manglem Singh

Professor CSE Dept.

NIT Manipur

Co-Investigator's Biodata

A. Name & Designation:	Dr. Yambem Jina Chanu Assistant Professor
B. Institution:	National Institute of Technology Manipur
C. Date of Birth:	15-05-1985
D. Whether belongs to SC/ST/OBC:	General
E. Academic & Professional career: **Academic career:**	9 years 2 months
F. Title of Doctoral thesis:	Development of new techniques for steganography and steganalysis
G. Award/Prize/Certificate etc. won by the Investigator:	NIL
H. Publications:	
Research Papers:	15 numbers

Dr. Yambem Jina Chanu

Asst. Prof. CSE Dept.

NIT Manipur

a) List of completed and on-going projects during the last five years (if any): NIL

Sl.No.	Title of the Project	Duration From To	Total Cost	Funding Agency

b) Project submitted for funding (if any): NIL

Sl.No.	Title of the Project	Name of Organization applied to	Funding Agency

2. Deviations made from original objectives, if any, while implementing the project, and reasons thereof :

No

3. Details of the project work. This should include full details of the Experimental set-up, Methodology adopted; Materials and Data collected and examined Data collection format/questionnaire etc. It should be supported by tables, charts, drawing, maps, photographs etc.

3.1Methodology:

The proposed system can be divided into the following parts:

3.1.1. Pre-processing:

Pre-processing of a signal can be said as applying any required form of processing to the signal in time domain before the feature extraction phase Normally, in the pre-processing stage the speech signal undergoes several common processes including analog to digital conversion (A/D) and enhancement, pre-emphasis filtering. The A/D process converts a sound pressure wave into its digital form. There are three steps in the A/D conversion process which is sampling, quantization and coding. The final product of this process is a digital version of the speech signal that can be processed by a computer. The pre emphasis filter is used to emphasis the speech spectrum above 1 kHz which contains important aspects of the speech signal and equalizes the speech propagation through air [45].

3.1.2. Feature Extraction

The MFCCs is perhaps the most popular and common feature of Speech Recognition System. Figure 4 shows the block diagram of the MFCCs. To obtain the MFCCs of a speech signal, the signal is first subjected to pre-emphasis filtering with the following finite impulse response (FIR) filter given by Equation (1.1)

$$H_{pre}(z) = \sum_{k=0} Na_{pre}(k) z^{-k} \quad (1.1)$$

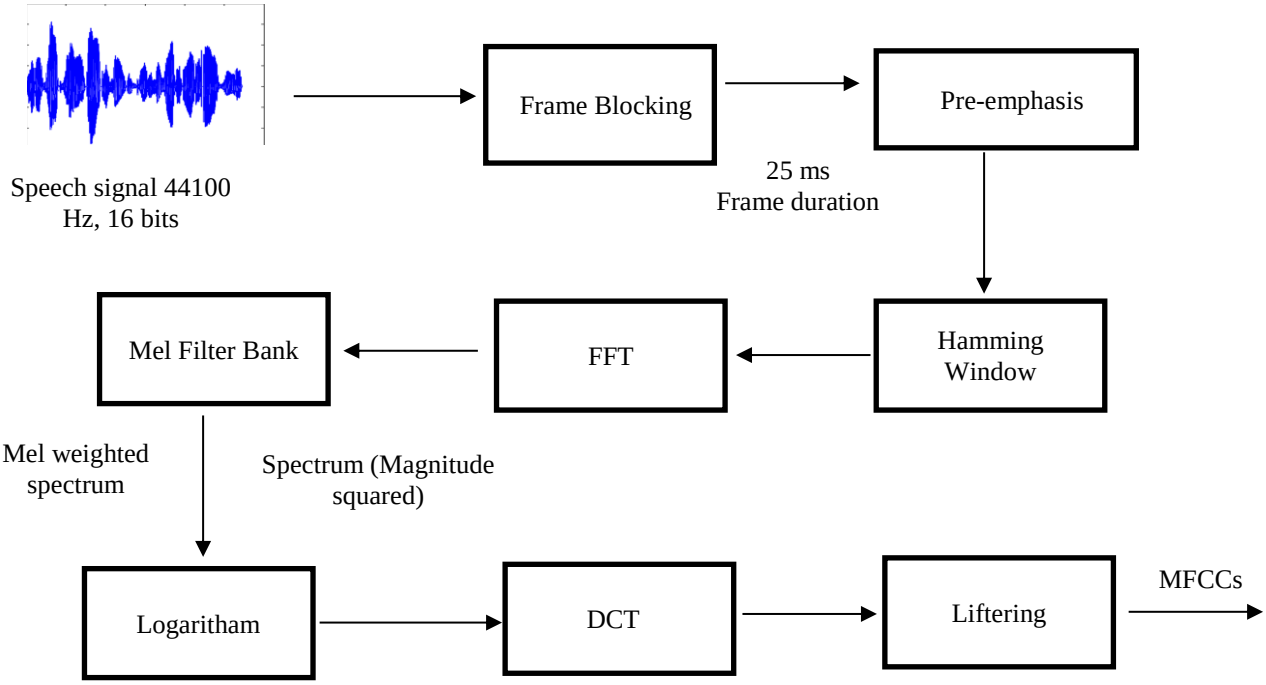


Figure 4: shows the block diagram of the MICC.

The value of the coefficient are usually takes the value between -1.0 to -0.4. However, in speech recognition systems values that are almost -1.0 are usually used [47]. The speech is processed on a frame-by-frame basis in what is called framing. Frame size of 25ms is used and windowing of these frames are to compensate discontinuities with the speech signal as a result of segmentation and overlapped frames. A hamming window is used by equation 1.2

$$w(n) = 0.54 + 0.46\cos(\frac{2\pi t}{T}) \quad (1.2)$$

Windowing means multiplying the window function w*(n)* with the framed speech signals *(n)* to obtained the windowed speech signal $S_{0w}(n)$ is given by equation (1.3):

$$S_{0w}(n) = s(n)w(n) \quad (1.3)$$

The discrete Fourier transform (DFT) of the windowed speech signal is then compared by equation (1.4):

$$S_{0w}(k) = \sum_{n=0}^{N-1} S_{0w}(n) e^{\frac{-j2\pi kn}{N}} \quad (1.4)$$

The mel- filterbank is a triangular bandpass filter which is equally spaced around the Mel-Scale. A Mel is a unit if perceived pitch or frequency of a tone. The mapping between real frequency (Hz) and Mel frequency is given by the equation (1.5):

$$fmel = 2595.\ log(1 + \frac{f}{700}) \quad (1.5)$$

The log of the mel frequency is taken. This step is to smooth unwanted ripples in the spectrum and done by the Equation (1.6)

$$mk = log\ fmel \quad (1.6)$$

Finally, DCT is applied to the log mel-cpestrum m_k as in Equation (1.6) to obtain the Mel-frequency Cepstral Coefficients (MFCC) c_i of the i^{th} frame given by Equation (1.7):

$$c_i = \sqrt{\frac{2}{N}} \sum_{k=1}^{n} m_k \cos\left(\frac{\pi i}{N}\ (k - 0.5)\right) \qquad (1.7)$$

3.1.3. Acoustic model:

In the proposed system DNN is used for acoustic modelling. A Deep layer perceptron with many hidden layers Neural Network is simply a multi-layers between its input and outputs. DNN with 3 hidden layer is used to c acoustic observation x into one of the context dependent phonetic states s. It is a nonlinear classifier that can be interpreted as a stack of log linear hidden layer models the posterior probabilities of a set of binary h given the input visible variables v, while the output layer model of the class posterior probabilities. Thus, in each of the hidden layers, the posterior distribution can be expressed as

$$p(hl|vl) = \prod_{j=1}^{N^l} p\left(h_{lj}\middle|v_l\right), \quad 0 \le l < L \qquad (1.8)$$

where,

$$p\left(h_{lj}\middle|v_l\right) = 1\frac{1}{1+e^{-zl,j(vl)}} z_{lj=W_{l,j}^T+b_{lj}} \qquad (1.9)$$

Each observation is propagated forward through the network, starting with the lowest layer (Vx). The output variables of each layer become the input variables of the next layer ie. v_1+1=h_1. In the final layer, the class posterior probabilities are computed using a softmax layer, defined as

$$p(s|x) = p(s|v_L) = \frac{e^{(zL,s(vL))}}{\sum_s e^{(zL,s(vL))}} \qquad (1.10)$$

In this work, networks are trained by maximizing the log posterior probability over the training examples, which is equivalent to minimizing the cross-entropy.

$$L = \sum_t p\ (s_l|x_l) \qquad (1.11)$$

The objective function is maximized using error back propagation which performs an efficient gradient-based update

$$\left(w_{l,j}, b_{l,j}\right) \leftarrow \left(w_{l,j}, b_{l,j}\right) + \eta\frac{\delta L}{\delta(w_{l,j},b_{l,j})}, \forall\ l, j \qquad (1.12)$$

Where, ɳ is the learning rate.

3.1.4. Language model:

Models that assign probabilities to sequences of words are called language models or LMs. In this chapter we introduce the simplest model that assigns probabilities to sentences and sequences of words, the N-gram. An N-gram is a sequence of N words: a 2-gram (or bigram) is a two-word sequence of words like "please turn, "turn your", or "your homework", and a 3-gram (or trigram) is a three-word sequence of words like "please turn your', or "turn your homework". N-gram is a conditional probability that if we observe some sequence of n - 1 words w_n^{n-1} then some particular nth word w^n will follow. It can be computed as the frequency of the word sequence w, also called w_1^n count and in further text denoted as C(wl n), divided by the frequency of the word sequence w-1. *N*gram language model is a collection of conditional probabilities for all n-word long sequences that can be composed of a given vocabulary. In languages with some relatively small vocabulary like Manipuri also probabilities of word triples can be computed and these are called trigrams. N-gram also called statistical language models are

successfully employed in tasks where spontaneous speech is recognized because they are robust and flexible in comparison to rule-based models that use some strictly described grammar

Maximum likelihood estimate (MLE) is the simplest possible n-gram model. It contains conditional probabilities for all possible n-grams as they appeared in the training data. Word sequences not present in the training corpus have their probabilities equal to 0. In a more formal way, the values in the model are computed according to the equation

$$p(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})} \quad (1.13)$$

3.1.5. Decoder:

The main task of decoder is to find the best word sequence for a given sequence of observation vectors which corresponds to Equation (1.3). This equation directly represents a search problem. The searching is done in the lattice. A lattice is a graphical structure which is used to represent the pruned hypothesis space which results after running recognition pass over the acoustic data. The lattice normally includes the hypothesized label (word and/or phone), the corresponding start and end times of the individual linguistic units, as well as the acoustic and language model likelihoods. Structurally, a lattice is essentially a directed, acyclic graph consisting of nodes and edges. Each edge is labelled with the hypothesized linguistic unit, the start and end times of that unit, and the associated acoustic and language model likelihood scores. Each node in the network corresponds to a point in time within the utterance. The identity of the start and end node for each edge is therefore used to impose the required structure.

3.2 System Implementation:

This section describes the implementation of the proposed ASR system. The next section explains data collection and preparation and the remaining section explains the implementation of ASR system.

3.2.1 Data Collection and Preparation:

As there are no database available for Manipuri dialects namely Kakching and Andro dialects. We collect the database from the native speakers of Kakching and Andro dialects. Speech recording is done with sampling frequency of 16 kHz and bit resolution of 16 bit in monochannel using wavesurfer. Native speakers of Kakching and Andro dialects were asked to read prompted sentences from Manipuri literature book. The resulting corpus consists of S hours of data by 60 native speakers, from across Kakching and Andro town in Manipur. Each Speaker read between 50 to 200 utterances. In total, corpus contains 3213 utterances. All speech data were recorded with a microphone in a quiet environment. This data has been used for training and testing the models. Data preparation is an important step in order to recognize speech in Kaldi. The following documents must be prepared as they are required for feature extraction, acoustic modelling and language modelling:

i. text: This file contains mappings between utterances and utterance ids which will be used by Kaldi

ii. spk2utt: This is a mapping between the speaker identifiers and all the utterance identifiers associated with the speaker

iii. utt2spk - This is a one-to-one mapping between utterance ids and the corresponding speaker identifiers.

iv. wav.scp - This file is actually read directly by Kaldi programs when doing feature extraction.

3.2.2 ASR system development details:

In recent years, ASR systems have become much more accurate and robust thanks to deep neural networks (DNNs). This research used scripts provided with Kaldi toolkit [49] for training DNN-based ASR systems, and IRSTI.M tool for building language models. Kaldi is based upon finite state transducers, and it is compiled against the OpenFst toolkit.

3.2.2.1 Acoustic modelling:

Figurc 5 shows the flow of the steps followed for training acoustic.

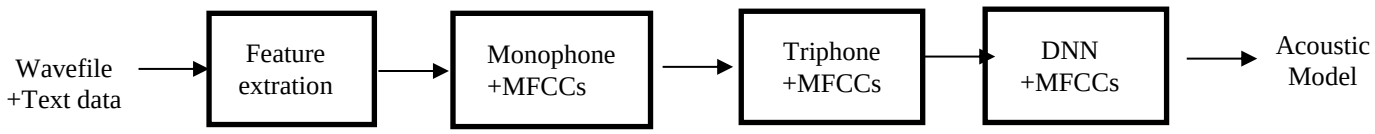


Figure 5: Steps of acoustic modelling.

Feature extraction: First, 13 dimensional Mel frequency cepstral coefficients (MFCCs) were extracted. Hamming window of 25 ms frame size and 10 ms frame shift was used. The MFCC statistics are computed for each frame.

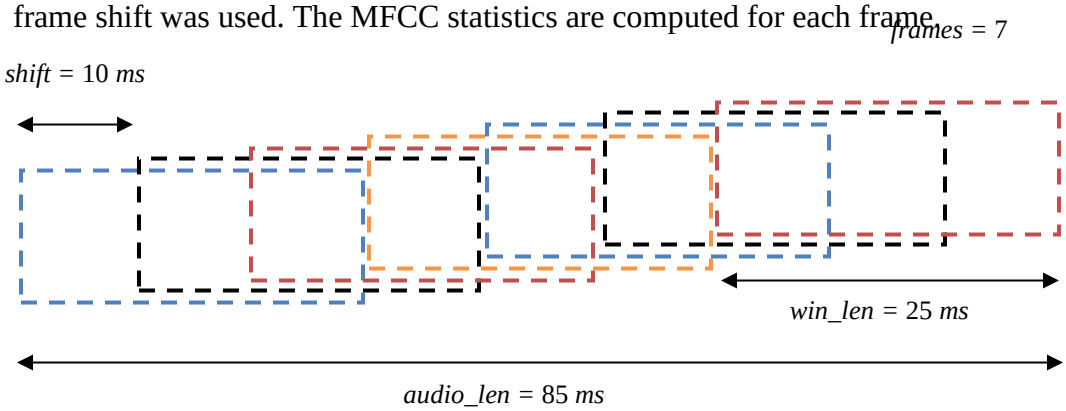


Figure 6: Frame shifting for feature extraction

In Figure 6 there are 7 windows frames with length of 25 ms and frame shift of 10 ms. The whole utterance lasts 85 ms.

DNN based acoustic model: Acoustic modelling by DNN architecture is shown in Figure 7. DNN has 3 hidden layer which takes the speech feature vectors as input and hidden layer uses tanh nonlinear activation function with softmax function in the output layer and gives the conditional probability of the feature vector for each phoneme class.

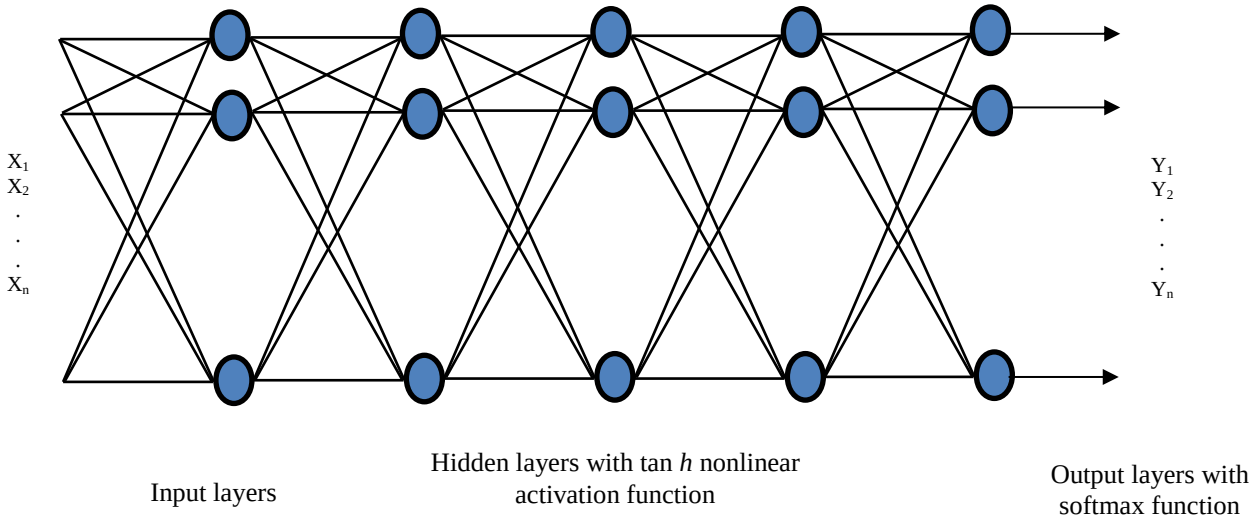


Figure 7: Framework of DNN based acoustic model.

3.2.2.2 Lexicon:

Lexicon is prepared from most frequent 932 words present in the text corpus of Manipuri dialects. The model of lexicon for few words is given in Figure 8.

3.2.2.3. Language modelling:

Language models are used to guide the search correct word sequence by predicting the likelihood of nth word using (n-1) preceding words. Language models can be classified into:

i. Uniform model: each word has equal probability of occurrence.

ii. Stochastic model: probability of occurrence of a word depends on the word preceding it.

iii. Finite state languages: languages use a finite3 state network to define the allowed word sequences.

iv. Context free grammar: It can be used to encode which kind of sentences is allowed.

We use the IRSTLM toolkit in Kaldi for training language models. Modified Kneser-Ney smoothing is used.

```
AcaNbsiH axx c a n b s i ng
AcOb axx c au b
AcuMbagidmQ a c u m b a g i d m k
AdoMd axx d o m d
AdoMg axx d o m g
AdoMgi axx d o m g i
AdoMgidmQ axx d o m g i d m k
Adu axx d u
Adubu axx d u b u
Adud axx d u d
Adudgi axx d u d g i
Adug axx d u g
AduguMb axx d u g u m b
AdumQ axx d u m k
Adun axx d u n
AE ai
Ae e
AEbu ai b u
AEbudi ai b u d i
AEdi ai d i
AEg ai g
AEga ai g a
AEgi ai g i
AEhaQ ai h a k
AEhaQki ai h a k k i
AEhaQsu ai h a k s u
AEKoI ai kh o i
AEKoIbu ai kh o i b u
```

Figure 8: Lexicon formats

3.2.2.4. Decoding:

First, hypothesis transcriptions were produced using the spliced MFCC features. Decoding of the audiobooks was done in two passes. In the first pass, a relatively computationally inexpensive language model was used to generate a lattice or word-graph of competing alternative hypotheses. In the second pass, a computationally expensive and higher order language model was used to re-score model probabilities on the lattice, re-rank the alternative hypotheses and find the 1-best hypothesis.

4. Outcome of the Project :

An Automatic Speech Recognition(ASR) system which can recognize Kakching and Andro dialects.

5. Analysis of Results:

In this section, two baseline systems are discussed which will be used for comparing results with the proposed system.

i. HMM-GMM with manophone training: A set of contexl-independent or monophone Gaussian mixture model-hidden Markov model (GMM-HMM) acoustic models is trained with MFCC features. The WER is calculated for this system and it is used for comparison with the proposed model

ii. HMM-GMM with triphone training: A set of context dependent phone i.e., phone model where every possible pair of left and right neighbours Gaussian mixture model-hidden Markov model (GMM-HMM) acoustic models is trained with MFCC features. The resulting models are called triphones. Context dependent phone modelling for Manipuri word is given in Figure 4.1. The WER is calculated for this system and it is used for comparison with the proposed model.

Evaluation Metric

The most common measures to evaluate the performance of speech recognition system are correctness, accuracy and error. These measures can be calculated on phoneme and word level. There are three types of mistake a speech recognition system can make:

i. Substitution: At the position of a unit (word or phoneme), a different unit has been recognized.

ii. Deletion: In the recognition result a unit is omitted.

iii. Insertion: The result contains additional units than were not spoken.

Correctness is defined as:

$$\text{Corr} = \frac{N-D-S}{N} \quad (1.14)$$

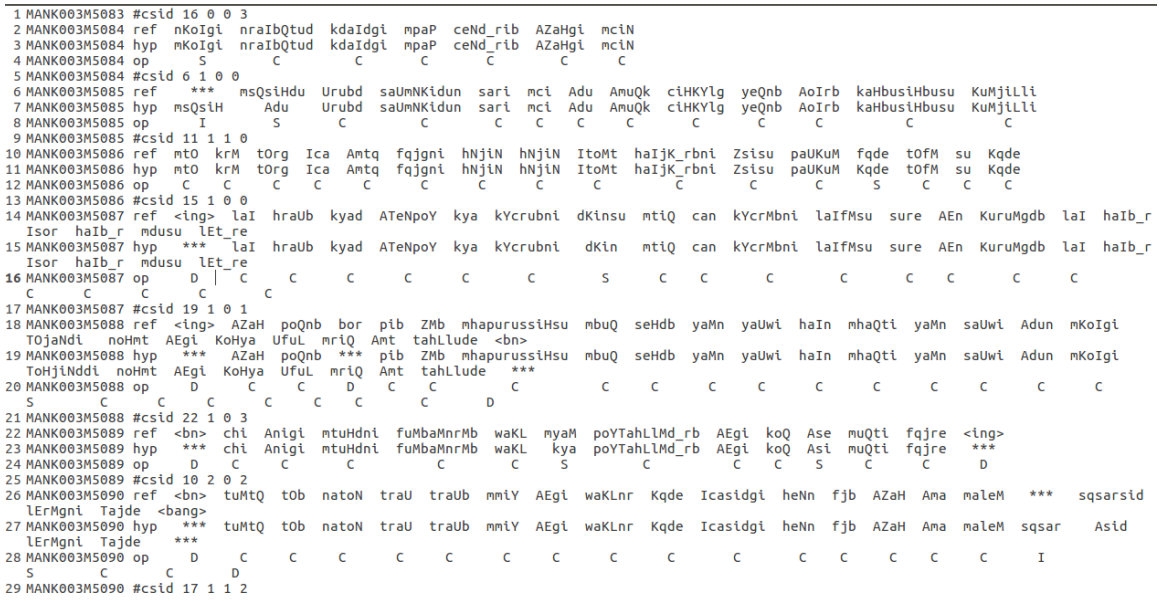
with N being the number of units in the correct transcription, D the number of deletions and S the number of substitutions. The accuracy is similar to the correctness, but additionally takes the number of insertions I into account:

$$\text{Acc} = \frac{N-D-S-I}{N} \quad (1.15)$$

Both measures are commonly given in percentage notation. Finally the error rate is the sum of all types of errors divided by number of words in the transcription. The error rate is closely related to the accuracy, usually only one of the two numbers is given.

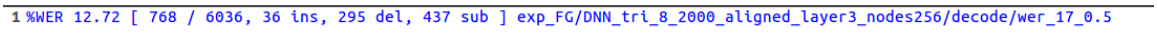
$$\text{WER} = \frac{D+S+I}{N} * 100 \quad (1.16)$$

Results:



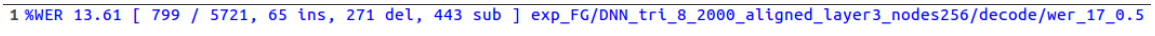
```
MANK003M5083 #csid 16 0 0 3
MANK003M5084 ref  nKoIgi  nraIbQtud  kdaIdgi  mpaP  ceNd_rib  AZaHgi  mciN
MANK003M5084 hyp  mKoIgi  nraIbQtud  kdaIdgi  mpaP  ceNd_rib  AZaHgi  mciN
MANK003M5084 op      S         C         C       C        C       C      C
MANK003M5084 #csid 6 1 0 0
MANK003M5085 ref    ***   msQsiHdu  Urubd  saUmNKidun  sari  mci  Adu  AmuQk  ciHKYlg  yeQnb  AoIrb  kaHbusiHbusu  KuMjiLli
MANK003M5085 hyp  msQsiH     Adu    Urubd  saUmNKidun  sari  mci  Adu  AmuQk  ciHKYlg  yeQnb  AoIrb  kaHbusiHbusu  KuMjiLli
MANK003M5085 op      I        S       C         C        C    C    C     C        C      C      C           C           C
MANK003M5085 #csid 11 1 1 0
MANK003M5086 ref  mtO  krM  tOrg  Ica  Amtq  fqjgni  hNjiN  hNjiN  ItoMt  haIjK_rbni  Zsisu  paUKuM  fqde  tOfM  su  Kqde
MANK003M5086 hyp  mtO  krM  tOrg  Ica  Amtq  fqjgni  hNjiN  hNjiN  ItoMt  haIjK_rbni  Zsisu  paUKuM  Kqde  tOfM  su  Kqde
MANK003M5086 op    C    C     C    C     C      C      C      C      C          C        C       C     S     C    C    C
MANK003M5086 #csid 15 1 0 0
MANK003M5087 ref  <ing>  laI  hraUb  kyad  ATeNpoY  kya  kYcrubni  dKinsu  mtiQ  can  kYcrMbni  laIfMsu  sure  AEn  KuruMgdb  laI  haIb_r
  Isor  haIb_r  mdusu  lEt_re
MANK003M5087 hyp   ***   laI  hraUb  kyad  ATeNpoY  kya  kYcrubni   dKin   mtiQ  can  kYcrMbni  laIfMsu  sure  AEn  KuruMgdb  laI  haIb_r
  Isor  haIb_r  mdusu  lEt_re
MANK003M5087 op     D     C     C      C      C      C       C        S      C    C       C        C       C    C       C      C
  C       C      C      C       C
MANK003M5087 #csid 19 1 0 1
MANK003M5088 ref  <ing>  AZaH  poQnb  bor  pib  ZMb  mhapurussiHsu  mbuQ  seHdb  yaMn  yaUwi  haIn  mhaQti  yaMn  saUwi  Adun  mKoIgi
  TOjaNdi   noHmt  AEgi  KoHya  UfuL  mriQ  Amt  tahLlude  <bn>
MANK003M5088 hyp   ***   AZaH  poQnb  ***  pib  ZMb  mhapurussiHsu  mbuQ  seHdb  yaMn  yaUwi  haIn  mhaQti  yaMn  saUwi  Adun  mKoIgi
  ToHjiNddi  noHmt  AEgi  KoHya  UfuL  mriQ  Amt  tahLlude   ***
MANK003M5088 op     D      C     D     C      C       C         C          C     C      C     C      C      C      C      C      C
  S       C      C      C      C     C      C       C       D
MANK003M5088 #csid 22 1 0 3
MANK003M5089 ref  <bn>  chi  Anigi  mtuHdni  fuMbaMnrMb  waKL  myaM  poYTahLlMd_rb  AEgi  koQ  Ase  muQti  fqjre  <ing>
MANK003M5089 hyp   ***  chi  Anigi  mtuHdni  fuMbaMnrMb  waKL   kya  poYTahLlMd_rb  AEgi  koQ  Asi  muQti  fqjre   ***
MANK003M5089 op     D    C     C        C          C       C     S          C          C    C    S     C      C      D
MANK003M5089 #csid 10 2 0 2
MANK003M5090 ref  <bn>  tuMtQ  tOb  natoN  traU  traUb  mmiY  AEgi  waKLnr  Kqde  Icasidgi  heNn  fjb  AZaH  Ama  maleM   ***   sqsarsid
  lErMgni  Tajde  <bang>
MANK003M5090 hyp   ***  tuMtQ  tOb  natoN  traU  traUb  mmiY  AEgi  waKLnr  Kqde  Icasidgi  heNn  fjb  AZaH  Ama  maleM  sqsar    Asid
  lErMgni  Tajde    ***
MANK003M5090 op     D     C     C     C      C     C      C     C      C      C       C       C    C     C     C     C      I
  S        C       C      D
MANK003M5090 #csid 17 1 1 2
```

Figure 9: Speech recognition output for DNN based acoustic modelling for Manipuri dialect



```
%WER 12.72 [ 768 / 6036, 36 ins, 295 del, 437 sub ] exp_FG/DNN_tri_8_2000_aligned_layer3_nodes256/decode/wer_17_0.5
```

Figure 10: WER DNN based acoustic modelling for Kakching dialect.



```
%WER 13.61 [ 799 / 5721, 65 ins, 271 del, 443 sub ] exp_FG/DNN_tri_8_2000_aligned_layer3_nodes256/decode/wer_17_0.5
```

Figure 11: WER DNN based acoustic modelling for Andro dialect.

Table1.1: WER comparison for DNN based model and baseline system for Kakching dialect.

ASR System	Total no. of words	Total no. of faults words	WER
HMM with monophone	6036	(79+319+513) = 911	15.09
HMM with triphone	6036	(41+362+497) = 900	14.91
DNN based system	6036	(36+295+437) = 768	12.72

Table 1.2: WER comparison for DNN based model and baseline system for Andro dialect

ASR System	Total no. of words	Total no. of faults words	WER
HMM with monophone	5721	(71+309+512) = 892	15.59
HMM with triphone	5721	(61+301+462) = 824	14.4
DNN based system	5721	(65+271+443) = 779	13.61

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
MANK005M	raw	20	313	261	21	2	31	54	20
MANK005M	sys	20	313	83.39	6.71	0.64	9.90	17.25	100.00
MANK006F	raw	103	1401	1234	97	9	70	176	81
MANK006F	sys	103	1401	88.08	6.92	0.64	5.00	12.56	78.64
MANK007F	raw	20	319	272	28	1	19	48	17
MANK007F	sys	20	319	85.27	8.78	0.31	5.96	15.05	85.00
MANK008M	raw	60	754	672	50	1	32	83	39
MANK008M	sys	60	754	89.12	6.63	0.13	4.24	11.01	65.00
MANK009F	raw	5	62	50	6	2	6	14	5
MANK009F	sys	5	62	80.65	9.68	3.23	9.68	22.58	100.00
MANK010M	raw	20	232	194	14	1	24	39	19
MANK010M	sys	20	232	83.62	6.03	0.43	10.34	16.81	95.00
MANKK011F	raw	60	754	670	52	3	32	87	39
MANKK011F	sys	60	754	88.86	6.90	0.40	4.24	11.54	65.00
MANKK012F	raw	40	562	510	31	2	23	56	5
MANKK012F	sys	40	562	90.74	5.51	0.35	4.09	9.96	100.00
MANKK013F	raw	45	603	559	23	2	21	46	19
MANKK013F	sys	45	603	92.70	3.81	0.33	3.48	7.62	95.00
SUM	raw	415	6036	5340	437	36	295	768	181
SUM	sys	415	6036	88.46	7.23	0.59	4.88	12.72	79.39

Figure 12: Number of words correctly recognized and faulty words per speaker for DNN based model for Manipuri dialect

Proposed ASR system achived 12.72% of WER has been obtained using DNN based acoustic modelling for Kakching Dialects and 13.61% of WER has been obtained using DNN based acoustic modelling for Andro Dialects.

Two Manipuri dialect namely Kakching and Andro dialect from their respective speaker is compare with respect to WER. Kakching dialect has lower WER (12.72%) then Andro dialect, WER (13.61). So, Kakching dialect higher speech recognized then Andro dialect.

6. Conclusion, summarizing the achievements and indicating scope of future work.

We presented a methodology for developing an ASR system that can recognize Manipuri dialects (Kakching and Andro). We are planning to refine and extend this work in the following ways:

i. Our research will continue on evaluating more features using various other feature extraction techniques to further improve the recognition.

ii. Other Manipuri dialects can also be incorporated.

iii. A relative study on various ANN models may be incorporated to conclude the best ANN.

iv. Differences in Manipuri dialects can also be further investigated.

7. The impact of different types of features, acoustic models and language models may be studied Benefits accorded from the Project :

a) Academic benefits: This project will be the base model for such research for future development.

b) Contributions towards socio-economic development: Effective and accurate transcription process will increase productivity and can enable too translate to other languages. And could develop many ASR for hospitals, Ima market etc.

References:

[1] Shahnawazuddin, S., et al. "Assamese spoken query system to access the price of agricultural commodities." *Communications (NCC), 2013 National Conference on*. IEEE, 2013.

[2] William J Frawley, "International encyclopedia of linguistics: AAVE-Esperato". Vol. 1, Oxford University Press, 2003, pp. 251.

[3] Pauthang Haokip, "The languages of Manipur A case Study of the Kuki –Chin languages", Linguistics of the Tibeto-Burman Area, Vol. 34.1, April 2011.

[4] PauthangHaokip, "Socio-linguistic Situation in Northeast India", Concept Publishing Company, 2011, p 61-62.

[5] http://meity.gov.in/content/language-computing-group-vi/accesed on 3/9/2017 accessed on 11/12/2018.

[6] Ossama Abdel-Hamid, et al. "Convolutional Neural Networks for Speech Recognition", IEEE/ACM TRANSCATION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 10, OCTOBER 2014.

[7] D Palaz, R Collobert, "Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks" - Proceedings of Interspeech, 2013 - infoscience.epfl.ch.

[8] X. Huang. A. Acero and H W. Hon (2001), Spoken Language Processing: a guide to theory, algorithm, and system development, Prentice Hall.. Hon (2001). Spoken Language Processing: a.

[9] D. Jurafsky and J. H. Martin (2009). Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall.

[10] R. Lawrence and B.-H. Juang (1993). Fundamentals of Speech Recognition Prentice-Hall, Inc.. (Engelwood, NJ).

[11] Enthnologue Languages of the world. Available from: URL https:/www.ethnologue.com/guides/how-many-languages accessed on August 2017.

[12] Pauthang Haokip (2011). The languages of Manipur A case Study of the Kuki-Chin languages. Linguistics of the Tibeto-Burman Area, vol.34.1.

[13] Pauthang Haokip (2011). Socio Linguistic Siuation in Northeast India Concept Publishing Company.

[14] Dr. Irom Robindro Singh (2013). The Status of Meiteilon among the Tibetd-Burman languages, Language in India available Url: Http:/languageinindia.com.vol. 13.9, pp. 320-332.

[15] William J Frawley (2003). International encyclopaedia of linguistics: AAVE-Esperato. Oxford University Press. vol. 1, p-251.

[16] Lawrence R. Rabiner and Ronald W. Schafer (2007), Introduction to Dgital Speech Processing; now Publishers Inc.

[17] D.Povey and A. Ghoshal 2011, The Kaldi Speech Recognition Toolkit, ASRU

[18] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri 2007. OpenFst: a general and efficient weighted finite-state transducer library, in Proceedings. CIAA.

[19] X. Huang and L. Deng (2010), An Overview of Modern Speech Recognition in Handbook of Natural Language Processing, Second Edition, Chapter 15, Chapman & Hall CRC, pp. 339-366.
[20] M. A. Anusuya and S. Katti (2011), Front end analysis of speech recognition: a review, Int. J. Speech Technology, vol. 14, no. 2, pp. 99-145.
[21] Kashyap Patel, R. K. Prasad (2013), Speech Recognition and Verification Using MFCC &VQ International Journal of Emerging Science and Engineering JESE, ISSN: 2319-6378, Volume-1, Issue-7,
[22] Oh-Wook Known, Kwokleung Chan, Te-Won Lee (2003), "Speech Feature Analysis Using Variational Bayesian PCA", in IEEE Signal Processing letters Vol. 10, pp.137-140.
[23] Thiang, and Wanto (2010), Speech Recognition Using LPC And HMM Applied for Controlling Movement of Mobile Robot, Seminar National le Informasi, pp. 67-72.od Teknoto.
[24] Onmesh Wadhwani, and Amit Kolthe (201), Recognition of Vermacoiur Language Speech for Discrete Words Using LPC Technique, Journal Research in Computer Science, Vol.2, No.9, pp.25-27 on of Vernacula
[25] Urmila Shrawankar, and Dr. Vilas The Extraction in Speech Recognition System: A Comparative Sludy kare (2010), Techniques for Feature Journal of Computer Applications in Engineering,. Technology and Science IJCAETS), ISSN 0974-3596, pp. 412-418.
[26] Vibha Tiwari (2010), MFCC and Its Applications in Speaker Recognition, International Journal on Emerging Technologies, Vol.1 Issue 1, pp. 19-22.
[27] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi (2010), Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Journal of Compuing, Vol.2 ISSUE 3, pp. 138-143
[28] M. A. Anusuya, S. K. Katti (2009), Speech Recognition by Machine: Review, aJCSis) International Journal of Computer Science and Information Securiy, Vol. 6, No. 3.
[29] Rajesh Kumar Aggarwal and M. Dave (2011), Acoustic modelling problem for automatic speech recognition system: advances and refinements Part (Part I), Int J Speech Technology, pp. 309-320.
[30] Dahl, G.E., Yu, D., Deng, L.,. Acero, A., (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang. Process. 20, pp. 30-42.
[31] Seide, F., Li, G., Chen, X., Yu, D., (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc.Workshop on Automatic Speech Recognition and Understanding, pp. 24-29.
[32] Hinton, G., Deng, L., Yu, D, Dahl, G., Mohamed, A., Jaitly., N., Senior. A.Vanhoucke, V., Nguyen, P., Sainath, T., Kinsgbury, B., (2012). Deep neural networks for acoustic modelling in speech recognition. IEEE Signal Process. Mag. 29, pp. 82-97.
[33] Morgan, N., Bourlard, H. (1995. Continuous speech recognition: an introduction to the hybrid HMM/ connectionist approach. IEEE Signal Process.Mag. 12, pp. 24-42.

[34] Hermansky, H., Ellis, D., Sharma, S., (2000). Tandem connectionist feature extraction for conventional HMM systems. In Signal Process, pp. 1635-1638.
[35] Nouza, J. (2000) Time Applications, Proc.3rd International Workshop on Text, Speech, Dialogue Springer-Verlag, Heidelberg, Germany, pp. 217-222
[36] R. Lawrence and BH. Juang (1993), Fandamentals of Speech Recognition. Prentice-Hall, Inc., (Engelwood, NJ).
[37] Ali. A.M.A., et al. (1998), An acoustic-phonetic feature-based system for automatic recognition of fricative consonants, Proc. IEEE ICASSP'98, pp961-964
[38] C.S.Myers and L.R.Rabiner (1981), A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition, EKE Tram. Acoustic Speech Signal Proc, ASSP-29:284-297.April 198
[39] D.R. Reddy (1996), An Approach to Computer speech Recognition by direct analysis of the speech wave, Tech Report No.C349, Computer Science Department Stanford University.
[40] H. Sakoe and S. Chiba (1978), Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustic Speech and Signal Processing, vol. 26, no. 1. pp. 43-49.
[41] J. K. Baker 1975, The Dragon System-An Overview, IEEE Trans on Acoustics Speech Signal Processing, Vol. ASSP-23, no. I, pp. 24-9,
[42] S. Xue O. A. Hamid H. Jiang L., Dai (2014), Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code Proc. of ICASSP, pp. 6339-6343.
[43] O. Abdel-Hamid A. Mohamed H. Jiang G. Penn (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, Proc. of ICASSP pp. 4277-4280.
[44] D. Palaz R. Collobert M. Magimai (2013), End to-end phoneme Sequence recognition using convolutional neural net-works, ArXiv e-prints.
[45] J. W. Picone (1993), Signal Modelling Techniques in Speech Recognition Proceedings of the IEEE, vol.81, pp. 1215-1247.
[46] D. O'Shaughnessy (2008), Invited Paper: Automatic Speech Recognition: History, Methods and Challenges, Pattern Recognition, vol, 41, pp. 28963-2979.
[47] V. Peddinti, D. Povey, and S. Khudanpur 2015, A time delay neural network architecture for efficient modelling of long temporal contexts, In Proc Interspeech.
[48] V. Peddinti, D. Povey, and S. Khudanpur (2015). A time delay neural network architecture for efficient modelling of long temporal contexts. In Proc, Interspeech.
[49] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. CGioel, Hannemann, P. Mot!l't cek, Y, Oian, P, Schwarz, et al. (2011) The Kaldi speech recognition toolkit.
[50] M. Federico, N, Bertoldi, and M. Cettolo (2008). IRSTLM: an open source toolkit for handling large scale language models. In Proc. Interspeech, pp. 1618-1621.
[51] C. Allauzen, M. Riley. J. Schalkwyk, W. Skut, and M. Mohri (2007), OpenFst: A general and efficient weighted finite-state transducer library. In Implementation and Application of Automata, pp. 11-23, Springer.

Annexure A

The primary goal of an ASR system is to hypothesize the most likely discrete symbol sequence out of all valid sequences in the language L, from the given acoustic input O [2]. As stated above, the input is treated set of discrete observations, such that:

$$O = (O_1, O_2, O_3, \dots, O_i) \quad \text{(A.1)}$$

Similarly, the symbol sequence to be recognized is defined as

$$W = (W_1, W_2, W_3, \dots, W_I) \quad \text{(A.2)}$$

The fundamental ASR system goal can be expressed as:

$$W = argmax\ P\ (W/O)\ for\ WCL \quad \text{(A.3)}$$

This equation implies that for a given sequence W and acoustic input sequence O, the probability P(W/O) need to be determined. Bayes' theorem can be applied to this probability to arrive at the following equation

$$P(W/O) = {}^{P\ (O/W)\ P(W)}\Big/_{P(O)} \quad \text{(A.4)}$$

The quantities on the right-hand side of the equation are easier to compute than P(W/O) P(W) is defined as the prior probability for the sequence itself. This is calculated by using prior knowledge of occurrences of the sequence W. Since the P(O) is the same for each candidate sentence W, thus equation can be simplified as
The probability P(O/W), which is the likelihood of the acoustic input O, the given the sequence W, is defined as the observation likelihood, can be called as acoustic score.