SPEAKER INDEPENDENT ISOLATED WORD SPEECH TO TEXT CONVERSION USING HTK

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Shweta Mittal
(801232026)

Under the supervision of:
Mr. Karun Verma
Assistant Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

July 2014

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Speaker Independent Isolated Word Speech To Text Conversion Using HTK*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Mr. Karun Verma* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Shweta Mittal

Signature:

(Shweta Mittal)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Mr. Karun Verma)
Assistant Professor,
CSED,
Thapar University,
Patiala

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

14/7

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude towards **Thapar University**, for providing me a platform to do my thesis work at such an esteemed institute.

I wish to express my respect, deep sense of gratitude and indebtedness to my guide **Mr. Karun Verma,** Assistant Professor, Computer Science and Engineering Department, Thapar University, Patiala for his invaluable and enthusiastic guidance, useful suggestions, unfailing patience and sustained encouragement throughout this work.

I would like to thank Dr. Deepak Garg, Head of Computer Science and Engineering Department, Thapar University, Patiala for kind help, guidance, encouragement and providing the necessary facilities to carry out my research. I am indebted to the faculty members of the department for valuable suggestions, friendly support and full cooperation rendered by all of them.

I am very grateful for the support I got from my family and friends. I would also like to express my gratitude to my parents for everything they have done for me.

Last, but not the least, I am thankful to Supreme Power "The God", one who has always guided me to work on the right path of the life. Without his grace, this would never come to be today's reality. With special thanks, I dedicate this thesis to God.

Shweta Mittal

ABSTRACT

Speech to Text Conversion or Speech Recognition allows a computer to identify the words that a person speaks into a mike or any other similar hardware and convert it into written words.

This thesis provides a description of implementation of HMM (Hidden Markov Model) Based Speaker Independent Isolated Word Speech to Text Conversion System. The System is developed by using HTK (Hidden Markov Model ToolKit) for Punjabi language which is an Indo-Aryan language spoken by about 130 million people mainly in West Punjab in Pakistan and in East Punjab in India. For implementation of the system, first of all, gathering of data that include 1010 words having 10 records from every 101 distinct words that is Punjabi language counting (0 to 100) is done from 10 distinct people. Then two sets of data are prepared in which first set of data obtained the above gathered data and second set of data obtained the above gathered data obtained after applying noise reduction technique (Auto Spectral Subtraction) on it. Then for both sets of data, out of 1010 words, 760 words are used to train the system and 250 words are used to test the system. The system uses the Mel Frequency Cepstral Coefficients (MFCCs) to extract features from speech files. For both sets of data, the system is trained and tested at three levels that are word level, mono-phone level and tri-phone level. The accuracy obtained from the system is 84.8% at word level, 88% at mono-phone level and 97.2% at tri-phone level for first set of data and 89.2% at word level, 92.4% at mono-phone level and 98.4% at tri-phone level for second set of data.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1
A BRIEF REVIEW OF SPEECH TO TEXT CONVERSION

1.1 Introduction to Speech to Text Conversion

Speech to Text Conversion or Speech Recognition allows a computer to identify the words that a person speaks into a mike or any other similar hardware and convert it into written words. Nowadays, a lot of research work is going on for the development of speech recognition systems with better accuracy. Many tools like KALDI, HTK, CMU SPHINX and others have been developed for the purpose of speech recognition. Speech to Text Conversion can be used in many applications like in mobile phones, computers, ATM machines, household appliances and others. It can help people with motor impairments that cannot use a standard keyboard and mouse to interact with computers more easily.

1.2 Basic Terminology

Few basic terms that are related to speech to text conversion and used in this thesis are:

1.2.1 Monophone

Monophone or phoneme is a basic unit of sound. Every word is composed of one or more monophones. For example, the word “tinn” is composed of three monophones that are “t”, “ih” and “n”. The Word “tinn” and monophones “t”, “ih” and “n” are transliterations of word and monophones of Punjabi word in English respectively.

1.2.2 Triphone

Grouping three phonemes and forming triphones as “P1-P2+P3” where the phoneme “P1” precedes phoneme “P2” and phoneme “P3” follows phoneme “P2” obtain Triphone. For example, the triphones “t+ih”, “t-ih+n” and “ih-n” are formed from three monophones “t”, “ih” and “n” that collectively form the word “tinn”.

1.2.3 Transliteration

Transliteration is text of source language written in target language which when pronounced produce the same sound as it is pronounced in source language. For example, the word “tinn” is transliteration of Punjabi word in English which when pronounced give the same sound as it gives when someone read the same word written in Punjabi.

1.2.4 Task Grammar

Task Grammar uses a set of rules to define a word so that the word can be easily recognized by the system. For the purpose, task grammar use some special symbols like braces to denote zero or more occurrences of word, square brackets to denote zero or one occurrence of word and others.

1.2.5 Task Dictionary

Task dictionary is a text file that defines monophones and labels corresponding to each word so that system can understand which label and phonemes are used for which word.

1.2.6 Vocabulary

Vocabulary defines the number of words used by the system for the purpose of training and testing of the system. Vocabulary can be small, medium, large and very large depending on the number of words used by the system. A small vocabulary can contain up to ten words, a medium can contain ten to hundreds of words, a large vocabulary can contain hundreds to thousands of words and a very large vocabulary can contain thousands to ten of thousands of words.

1.3 Types of Speech Recognition

On the basis of type of speakers for which system is developed, Speech Recognition may be

- Speaker Dependent
- Speaker Independent

In **Speaker Dependent Speech to Text conversion**, system is developed for particular group of persons. In **Speaker Independent Speech to Text conversion**,

system is developed for any kind of speaker. Speaker Dependent systems are more accurate as compared Speaker Independent systems but they are not as flexible as Speaker Independent systems as they are limited for a particular group of persons.

On the basis of way to recognize speech, Speech Recognition may be

- **Isolated Word Speech Recognition** in which system recognizes only one word at a time and word is preceded and followed by silence.
- **Connected Word Speech Recognition** in which system recognizes speech containing more than one words. Words are separated by small silences.
- **Continuous Word Speech Recognition** in which system recognizes speech containing more than one words and words are connected without any silence.
- **Spontaneous Word Speech Recognition** in which system recognizes speech in which not necessary words like “ums”, “ahs” and others are present along with words. Motive of Spontaneous Word Speech recognition is to recognize natural speech.

1.4 Basic Methodology used in Speech to Text Conversion

A basic Speech To Text Conversion System includes five steps to recognize a word that are Preprocessing, Feature Extraction, Acoustic Models, Language Model and Recognition shown in Figure 1.1.

1.4.1 Preprocessing

Preprocessing includes conversion of input speech file into a form that can be easily understandable by the system. For the purpose, input speech signal in the form of analog signal is converted into a digital speech signal. The digitized speech signal is then processed through low-pass filters for increasing the magnitude of higher frequencies with respect to the magnitude of lower frequencies. This process is known as preemphasis. The preemphasized speech is then converted into the frames of subsequent samples with frame size ranges from 10 to 25 milliseconds and an overlap of 50 to 70% between neighboring frames. Each individual frame is windowed to minimize the signal discontinuities at the beginning and at the end of each frame.

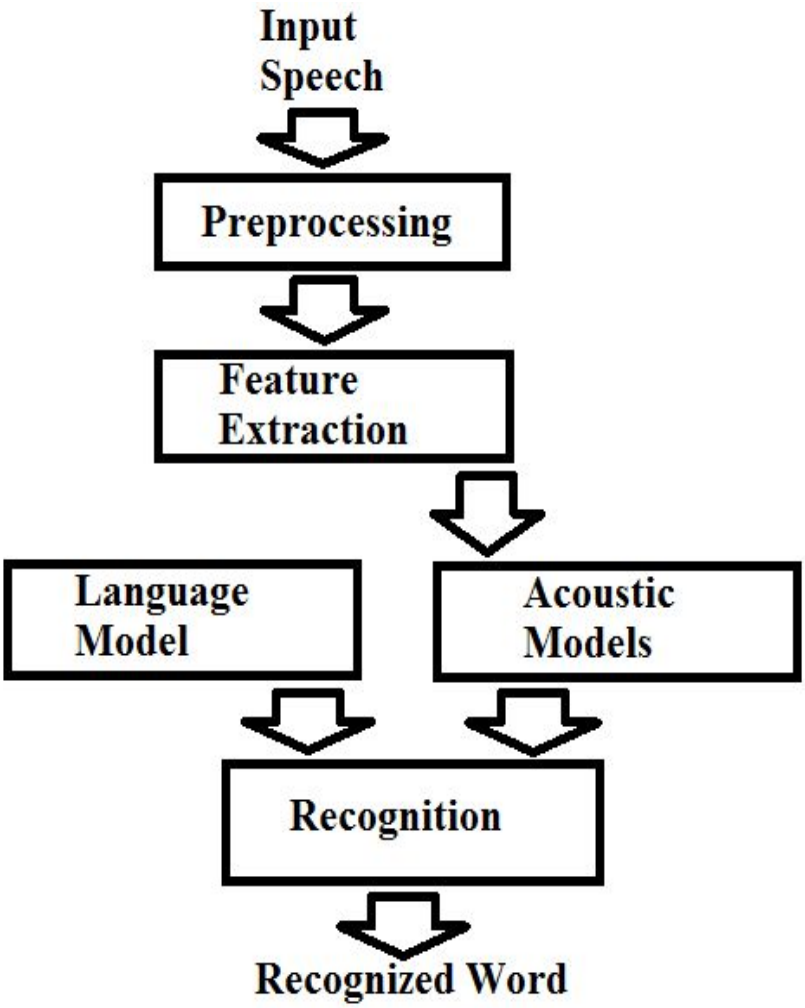


Figure 1.1: Basic Methodology used by Speech to Text Conversion System to recognize a word

1.4.2 Feature Extraction

It involves the extraction of useful information from the speech signal by parameterization of speech. The main purpose of feature extraction is to discard the irrelevant information from the speech signal. The capability of feature extraction is measured in terms of how it can find out actual speech signal information.

1.4.3 Acoustic Models

Acoustic Models are some reference models with which the speech signal is compared for the purpose of recognition. Acoustic Models are of two types. They may be word models or phoneme models. Different techniques like Hidden Markov Model (HMM), Support Vector Machine (SVM), Artificial Neural Network (ANN) and others can be used to generate the acoustic models.

1.4.4 Language Model

Language Model involves the creation of grammar that defines the set of rules to define a word so that the word can be easily recognized by the system. For the purpose, task grammar use some special symbols like braces, square brackets and others. Language Model defines the validity of word.

1.4.5 Recognition

The word corresponding to input speech signal is recognized with the help of Acoustic Models and Language Models.

1.5 Hidden Markov Model Toolkit (HTK)

HTK (Hidden Markov Model Toolkit) is an open source that is used for speech recognition and consists of a set of library modules and tools available in C source form [1]. It is developed by the CUED (Cambridge University Engineering Department).

There are multiple tools provided by HTK, which perform different functionalities. This report uses some of the HTK tools for the purpose of speech recognition.

Standard options of HTK Tools are:

- -A : used to display the command line arguments.
- -D : used to display configuration settings.
- -T 1 : used to display some information about the algorithm actions.

1.6 Introduction to HMM

Hidden Markov Model (HMM) is a weighted finite-state automata [2] which consists of a sequence of states q = (q0,q1,q2,…..qn) and a set of transition probabilities between states. An adjacency matrix can represent set of transition probabilities. According to the way, the sequence of states represent what, different levels can be defined. At word level, sequence of states represent one word that means HMM model is defined for one word. At mono-phone level, one HMM model is defined for each phoneme of the word. At tri-phone level, one HMM model is defined for each triphone consisting of group of three phonemes.

1.7 Noise Reduction

Noise reduction is very necessary to improve the accuracy of Speech to Text Conversion System, as noise is a common factor that degrades the quality of speech. The purpose of noise reduction is to lowers the level of noise without affecting the quality of speech.

1.7.1 Auto Spectral Subtraction

Auto Spectral Subtraction [3] is a simple and efficient noise reduction technique. In this technique, an average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) [4] is improved.

1.7.2 WavePad Sound Editor

WavePad Sound Editor is a tool used for noise reduction. The Editor uses Batch converter that allows transformations of many files at a time. Batch converter allows adding many files for the purpose of noise reduction. Batch converter has an option to add command to perform Auto Spectral Subtraction technique of noise reduction that allows the performance of Auto Spectral Subtraction technique on added speech files.

CHAPTER-2
LITERATURE REVIEW

A Spontaneous speech recognition system has been built using HTK for credit card corpus that was provided for summer workshop held at Rutgers CAIP Center [5]. The workshop was provided on Robust Speech Processing. The overall accuracy obtained from the system war very low due to some problems with credit card data.

For transcription of conversational telephone system, new features were integrated in CU-HTK (Cambridge University HTK) [6]. Many improvements were achieved and overall word error rate achieved was 25.4% that was the best performance by a statistically significant margin.

A Voice-controlled mobile robotic system was developed that was capable of controlling a robot by giving simple human voice commands [7]. The system was capable of relaying recognized voice commands from human to a mobile robot.

CU-HTK Mandarin BN (Broadcast News) transcription system was developed that was capable of handling both Mandarin and English data, as some English data was requirement of Mandarin Broadcast News Task [8].

A speech recognition system was developed to recognize speech in Polish language [9]. The system was trained by vocabulary of 365 words gathered from 26 males. The features of Polish were also described as they differ from English language.

A speaker-independent pronunciation recognition and assessment system was built using HTK by improving acoustic modeling [10]. The system used 673 words by using Chinese learning framework. The system allowed evaluation of quality of pronunciation by using results obtained by HTK and Viterbi coding.

Using Bengali speech corpus for two different age groups did speaker independent continuous speech recognition [11]. Two age groups were younger group having age

among 20 to 40 and older group having age among 60 to 80. For alignment of data, Hidden Markov Model Toolkit (HTK) was used. Speech corpus quality was checked by using performance of phoneme and continuous word recognition.

To reduce the time used by the training process of HTK (Hidden Markov Model Toolkit) for its high computations, paraTraining (a parallel training model) in HTK was designed [12]. Also various optimization methods were developed for improving performance of HTK on GPU.

An automatic speech recognition system has been developed for recognizing isolated and connected words of Punjabi language by using word model and triphone model on HTK 3.4.1 [13]. 200 words are used by the system for building vocabulary. The accuracy obtained by the system was 92.05% at word level and 97.14% at triphone level for isolated words and 87.75% at word level and 91.62% at triphone level for connected words.

As Speech is a natural form of communication, Microsoft Corporation developed Speech Application Program Interface (SAPI) for speech related works in its Windows operating systems that includes features for only eight languages including English [14]. They managed SAPI to match pronunciation from continuous Bangla speech in precompiled grammar file of SAPI and SAPI returned Bangla words in English character if matches occur. The words are then used to fetch Bangla words from database and return words in true Bangla characters and to complete the sentences.

The access to communication technologies has be-come essential for the handicapped people. This study introduces the initial step of an automatic translation system able to translate visual speech used by deaf individuals to text, or auditory speech [15]. Such a system would enable deaf users to communicate with each other and with normal-hearing people through telephone networks or through Internet by only using telephone devices equipped with simple cameras.

The problems in increasing the size of large vocabulary of speech recognition system that was built for clean as well as noisy read speech tasks were examined to handle broadcast news transcription [16]. The HTK system for H4 evaluation was described in which over pervious HTK large vocabulary systems, a number of new features were included.

Two most popular open source speech recognition systems, HTK and SPHINX were compared on Chinese Mandarin [17]. Both systems were used to recognize isolated words and continuous sentences. The performance of SPHINX3 was observed better than that of HTK. It was also observed that performance of system could be improved by using pitch information, as Mandarin is a tonal language.

Arabic speech recognition system was built using HTK [18]. First of all, Composition of words to its phones was done to build an Arabic dictionary. Then speech feature vectors were extracted by using Mel frequency Cepstral Coefficients. Then training of data was done to estimate the parameters of HMM. Then testing was done by using 10 speaker dependent and 3 speaker independent samples. Overall system performance obtained for sentence correction, word correction and word accuracy were 90.62%, 98.01% and 97.99% respectively.

Vietnamese Speech Recognition system was built using HTK [19]. Web was used for gathering data and CMU SLM (Carnegie Mellon Statistical Language Modeling) Toolkit was used for building bigram language. Different experiments were done using different acoustic models.

An isolated automatic speech recognition system was built using HTK [20]. First of all, gathering of 115 distinct Punjabi words was done from 8 persons. Then the system was trained for those 115 words. Then testing was done by gathering data from 6 persons. A GUI using JAVA was developed to make the system more interactive. Description of role of HTK tools for different phases of the system was also given.

To reduce the time that is needed to manually labeling of speech files, a package HTKTrain has been defined for the purpose of segmentation of speech [21]. To test

the package, its performance has been compared with labeling that has been done manually for a Hindi database and results are good showing average deviation of 28ms as compared to the labeling of Hindi database that has been done manually.

A Continuous Speech recognition system has been developed using HTK at two levels that are monophone level and triphone level [22]. The system uses vocabulary of 16800 words having 42 distinct words that include numerals zero to nine and names of 20 people. The results obtained are 74.11% at monophone level and 93.77% at triphone level.

Two Speech recognition systems have been developed using HTK. For the first system, the data has been prepared by taking data from South African people who use a variety of accents for the purpose of training of the system [23]. For the second system, the data has been prepared by taking data from people having English accent for the purpose of training of the system. When both systems were compared with each other, it was found that first system is slightly better than the second system as accuracy obtained during testing by the first system is better as compared to the accuracy obtained during testing by the second system. So it was observed that accent of people has major role in improving the accuracy of a speech recognition system.

A Hidden Markov Model has been defined for recognizing emotions from speech signals [24]. The system has used German database for the purpose of recognizing emotions from speech files. For the purpose, different characteristics of chosen emotions are compared with features that have been extracted from speech files.

A 2003 CU-HTK Large Vocabulary speech recognition system has been developed for recognizing CTS (Conversational Telephone Speech) and its performance for RT-03 (2003 Rich Transcription) was discussed [25]. A 2004 CU-HTK system has been developed and its performance has been discussed for RT-04 after that it was obtained that 2004 CU-HTK system showed 24% reduction in word error rate as compared to 2003 CU-HTK system [26]. Improvement was done on CU-HTK Broadcast News Transcription system that showed 25% reduction in word error rate as compared to 2003 CU-HTK system [27].

MFCCs (Mel Frequency Cepstral Coefficients) are one of the major features that are extracted from the speech signals. Optimization of these coefficients has been done [28] and development of a speech recognition system has been done for testing the optimization. On testing, it was observed that there was 12.02% improvement in the accuracy rate of recognition by the system.

A Speech recognition system has been developed for Telgu language [29]. The system used HTK for the development. The system was trained for continuous Telgu speech. The data was gathered from male persons for the purpose of training and testing the system.

Various issues of Speech to text conversion system have been discussed like overlapping of speech, low SNR (Signal to Noise Ratio), effectiveness and others [30].

An Automated Isolated Word Speech Recognition System has been developed for Hindi language. The System used HTK for the development [31]. The database is prepared by collecting data from nine people. The system used 113 Hindi words for the purpose of training. The 10 states HMMs were used by the system for the purpose of training. The accuracy obtained by the system was 95.49%.

CHAPTER-3
PROBLEM STATEMENT

Speech to Text Conversion or Speech Recognition is the process of converting spoken words into written text. It involves gathering of spoken words through mike or any other similar hardware and converting the words into written text. It can be used in many applications like in mobile phones, computers, ATM machines, household appliances and others. Nowadays, a lot of research work is going on for the development of speech recognition systems with better accuracy. Many tools like KALDI, HTK, CMU SPHINX and others have been developed for the purpose of speech recognition. Many speech recognition systems have been developed. Disturbances caused by environment and other causes, different styles of speaking used by humans are the major reasons that existing speech recognition systems are not 100% accurate. Accuracy of the system is the major concern.

Objective

Objective of this thesis is to give a minor contribution to sort out above stated problem to some extent by developing a system consisting of vocabulary of 1010 words having 101 distinct Punjabi words. The System is capable of isolated recognition of the words present in vocabulary at three levels that are word level, mono-phone level and tri-phone level. To improve the accuracy of the system, a noise reduction technique (Auto Spectral Subtraction) is applied at the time of data preparation of 1010 words.

CHAPTER-4
IMPLENTATION OF HMM-BASED PUNJABI SPEAKER INDEPENDENT ISOLATED WORD SPEECH TO TEXT CONVERSION USING HTK

4.1 Pre-Requisites

For implementation of the system, Pre-Requisites are

- Linux Environment like Ubuntu.
- Audacity tool for recording of speech files.
- WaveSurfer tool for labeling of speech files.
- WavePad Sound Editor for noise reduction in speech files.
- HTK tool for the purpose of speech recognition.

This thesis uses following versions of softwares for the implementation

- Ubuntu 10.04.4 32 bit
- Audacity 1.3.12-beta
- WaveSurfer 1.8.5
- WavePad Sound Editor
- HTK 3.4.1

4.2 Database Preparation

Database is prepared by gathering of data that include 1010 words having 10 records from every 101 distinct words that is Punjabi counting (0 to 100) is done from 10 distinct people. Database preparation is done in two ways that means two sets of data are prepared. In first set of data, Database preparation involves two steps that is recording of speech files and then labeling of that speech files. In second set of data, Database preparation involves three steps in which first two steps are same as that of first set and third step is noise reduction in speech files.

4.2.1 Recording and Labeling of Speech files

In the system, speech files are recorded with the help of Unidirectional Panasonic Karaoke Microphone and Audacity tool. The speech files are in wave format (.wav). With the help of Audacity tool, the sample frequency is set to 16 KHz. Figure 4.1 shows the waveform obtained for Punjabi word “ikk” with the help of tool Audacity.

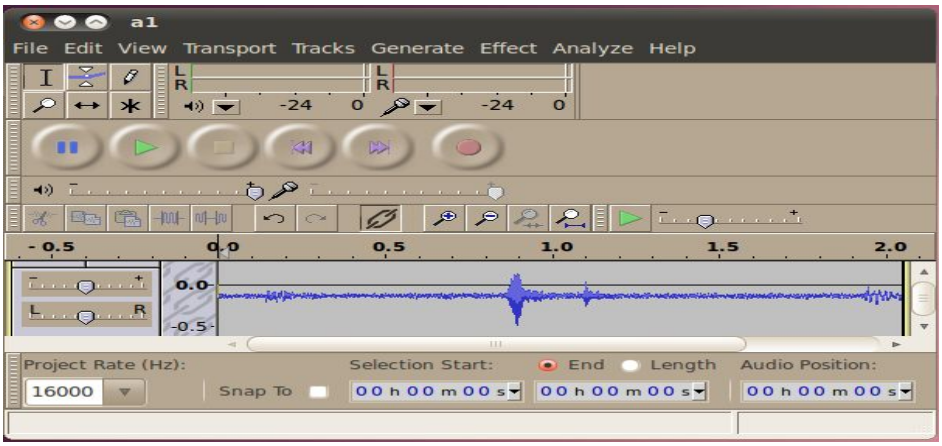


Figure 4.1: Waveform obtained for Punjabi word “ikk”

Labeling of speech files is done and .lab files are obtained with the help of tool WaveSurfer. Figure 4.2 shows HTK transcription of the word “ikk” obtained with the help of tool WaveSurfer.

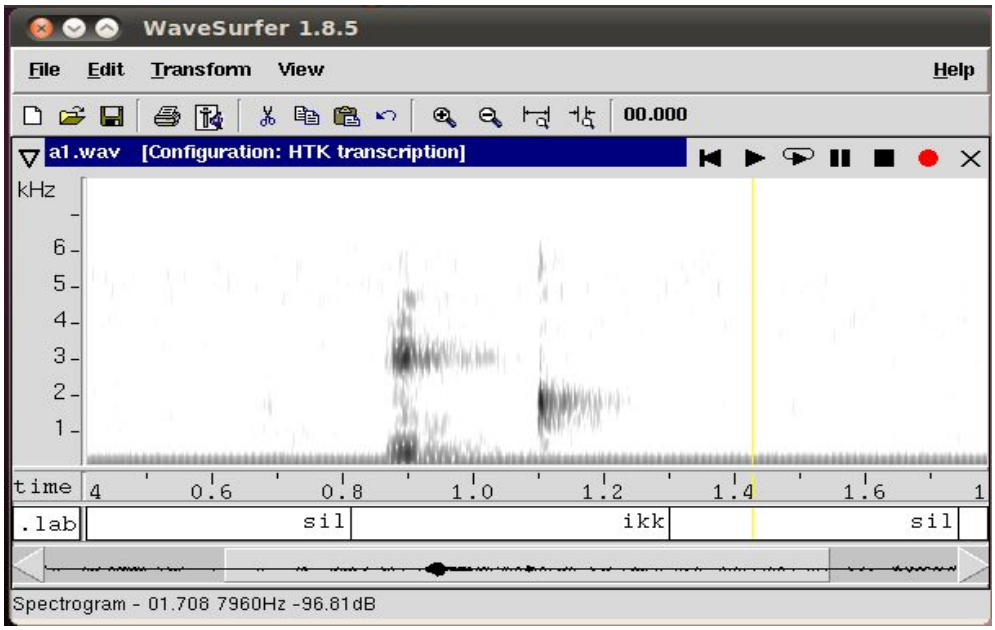


Figure 4.2: HTK transcription of Punjabi word “ikk”

For labeling of speech files, sil label is used to represent silence before and after the word whose labeling is done and word itself is labeled with the transliteration of that Punjabi word in English. The .lab files obtained with the help of tool WaveSurfer are the HTK transcription of the words.

4.2.2 Noise Reduction in Speech files

Noise reduction is done with the help of WavePad Sound Editor. Noise reduction does not affect quality of speech. It only lowers the level of noise. The system uses Batch converter that allows transformations of many files at a time. Figure 4.3 and Figure 4.4 shows how the system add many files to batch converter to perform noise reduction.

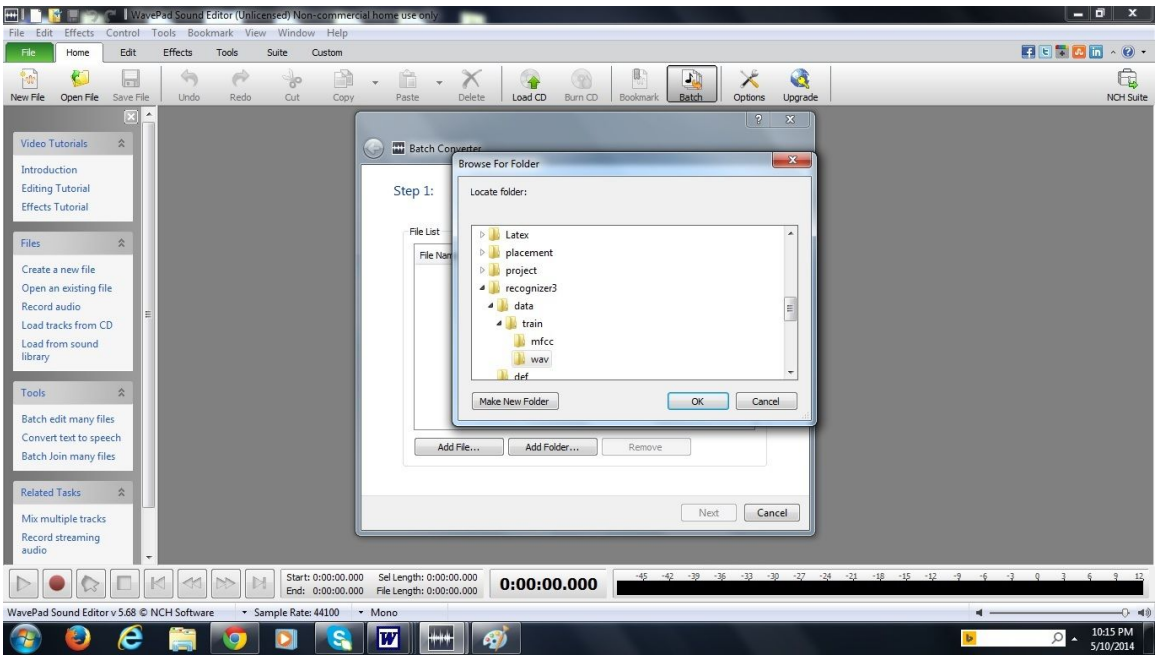


Figure 4.3: Add folder "wav" containing speech files used for Database Preparation to Batch Converter

After adding the speech files used for database preparation to Batch converter, the system adds command "Noise Reduction (Auto Spectral Subtraction)" to Batch converter. Figure 4.5 shows how the system adds command for noise reduction to Batch converter.

After that, the system performs noise reduction in speech files and saves the speech files with the same format (.wav) that is shown in Figure 4.6.

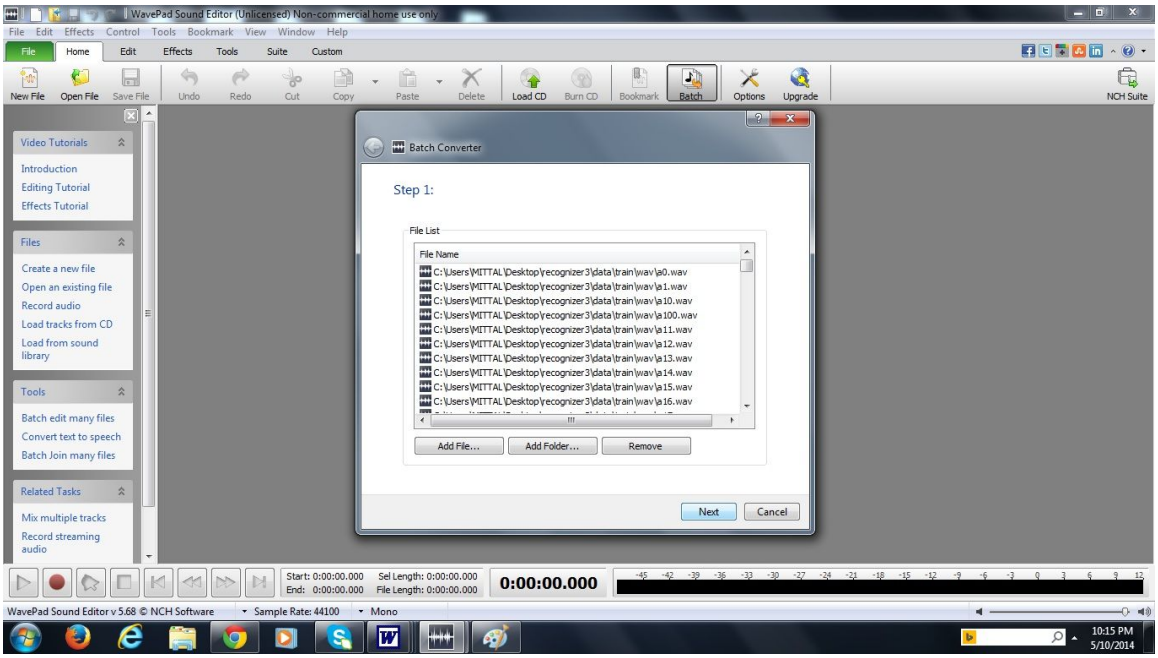


Figure 4.4: Batch Converter containing speech files used for Database Preparation

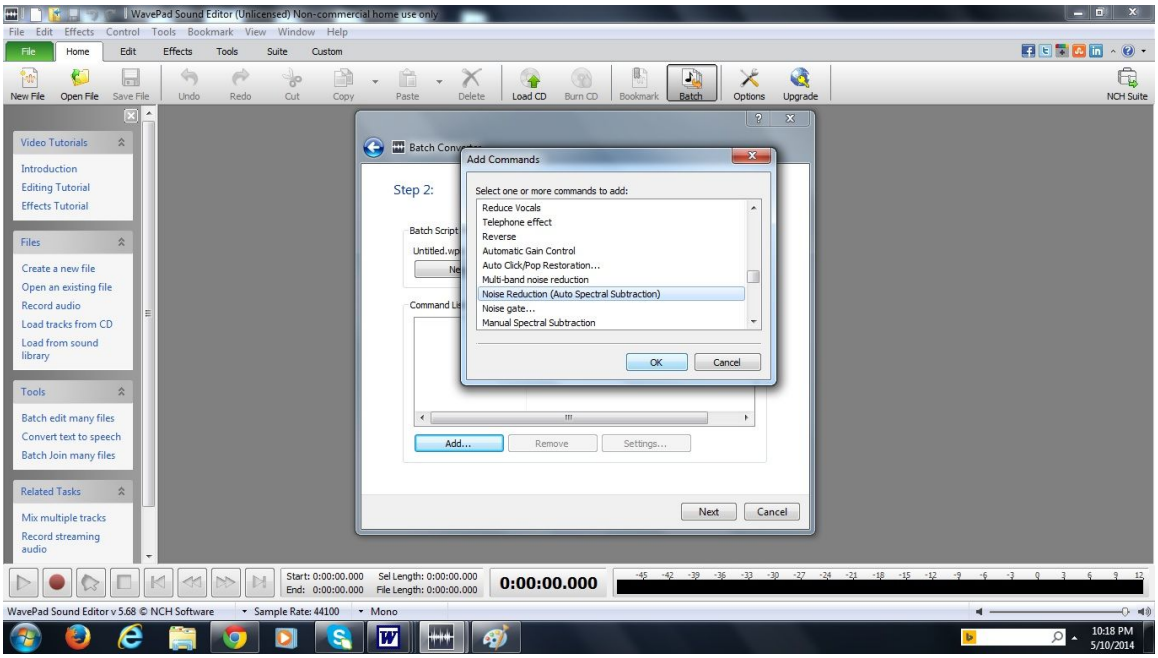


Figure 4.5: Adding of command "Noise Reduction (Auto Spectral Subtraction)" for noise reduction to Batch converter.

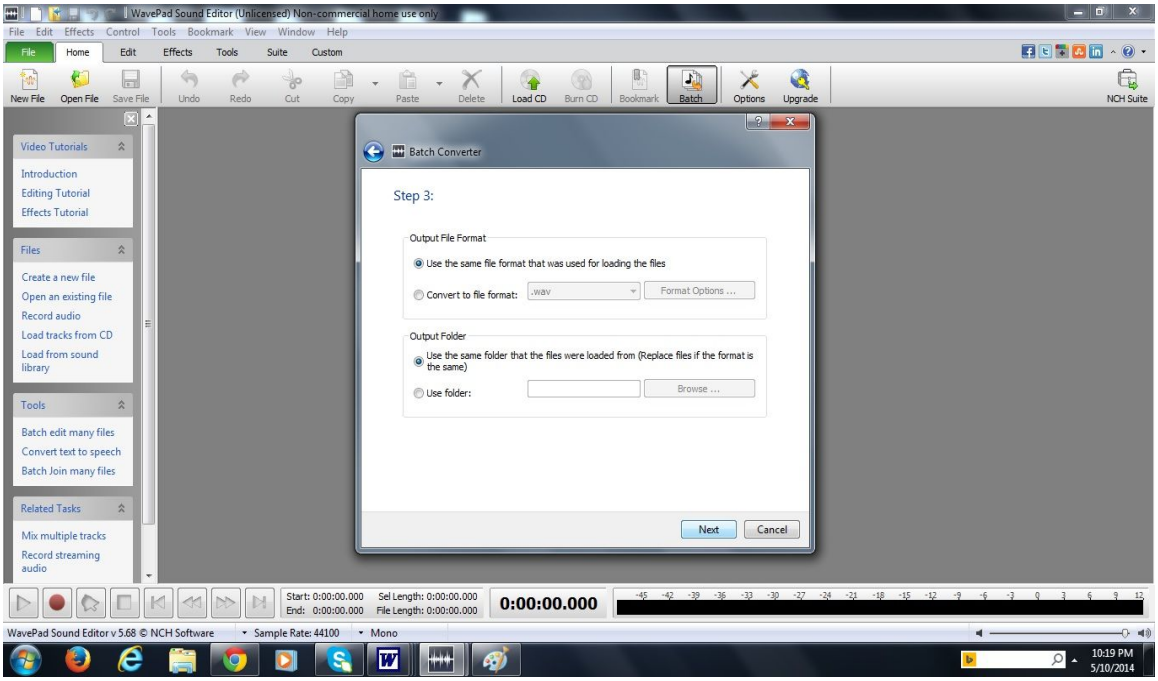


Figure 4.6: The System allows the process of noise reduction by allowing same format of speech files (.wav)

4.3 Task Definition

Task grammar is created in the system to define the words to be recognized by the system. In the grammar, the braces denote zero or more occurrences of START_SIL and END_SIL and square brackets denote zero or one occurrence of WORD. Also, Task dictionary is created by the system to define the label and HMM corresponding to each word defined in the grammar. Task dictionary is used in generation of transcriptions for monophones and triphones that are used in training and testing of the system. In task dictionary, the system use transliterations of the words and phonemes of Punjabi in English as shown in Table 4.1. After creating task grammar and task dictionary, compilation of the task grammar is done by the system with the help of HParse tool of HTK.

Task grammar used by the system is:

$WORD = CIPHER | IKK | DO | TINN | CHAR | PANJ | CHHE | SATT | ATTH | NAUM | DAS | GIARAM | BARAM | TERAM | CHAUDAM | PANDRAM | SOLAM | SATARAM | ATHARAM | UNNI | VIH | IKKI | BAI | TEI | CHAUVI | PACHI | CHHABBI | SATAI | ATHAI | UNTTI | TIH | IKTTI | BATTI | TETI | CHAUNTI | PAINTI | CHHATTI | SAINTI | ATHTTI | UNTALI | CHALI | IKTALI | BATALI | TARTALI | CHUTALI | PANTALI | CHHIALI | SANTALI | ATHTALI | UNANJA | PANJAH | IKVANJA | BAVANJA | TARVANJA | CHURANJA |

PACHVANJA | CHHAPANJA | SATVANJA | ATHVANJA | UNAHAT | SATTH | IKAHATH | BAHATH | TREHAT | CHAUNHAT | PEHANT | CHHEHAT | SATAHAT | ATAHAT | UNATTAR | SATTAR | IKHTTAR | BAHTTAR | TIHATTAR | CHUHATTAR | PANJHATTAR | CHHIHATTAR | SATATTAR | ATHATTAR | UNASI | ASSI | IKASI | BIASI | TARIASI | CHURASI | PACHASI | CHHIASI | SATASI | ATHASI | UNANVE | NABBE | IKANVEM | BANVEM | TIRIANAVAN | CHURRANAVAN | PACHANNAVEN | CHHIANNAVEN | SATANNAVEN | ATHANNAVEN | NARHINNAVEN | SAU;

({ START_SIL } [$WORD] { END_SIL })

4.4 Transcription Generation

The system generates three transcriptions for the three levels. Figure 4.7 shows how the three transcriptions define HMM model for word ”tinn” with words at word level, with monophones at mono-phone level and with triphones at tri-phone level.

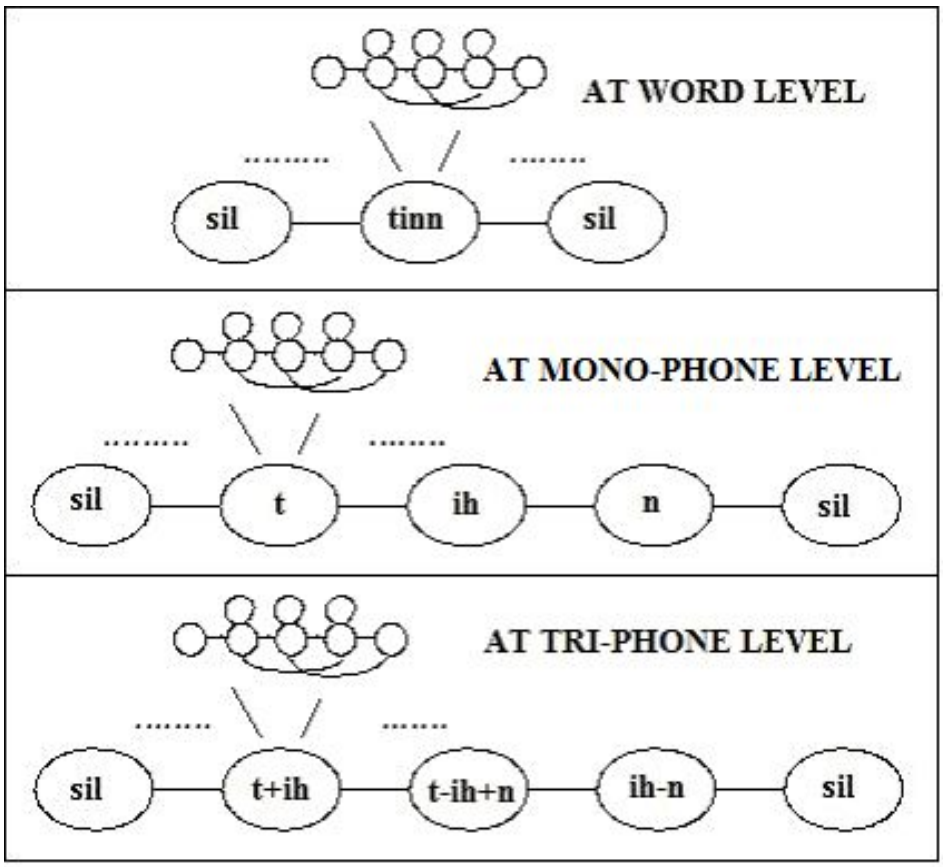


Figure 4.7: HMM model at the three levels for word “tinn”

4.4.1 Word Level Transcription

Word level Transcription named as Master label file (.mlf) [1] is created by concatenating the HTK transcriptions that are obtained at the time of database preparation using WaveSurfer tool.

4.4.2 Mono-Phone Level Transcription

In Mono-Phone level Transcription, Every word is expanded to number of phonemes. For the purpose, HDMan tool of HTK is used to obtain the phonemes corresponding to the words present in the task dictionary. The phonemes obtained by HDMan tool are used by HLEd tool of HTK to expand Word level transcription to Mono-phone level transcription. Table 4.1 shows the transliterations of the words and their corresponding phonemes used in the dictionary of the system.

4.4.3 Tri-Phone Level Transcription

In Tri-Phone level Transcription, triphones are obtained by grouping three phonemes and forming triphones as ”P1-P2+P3” where the phoneme ”P1” precedes phoneme ”P2” and phoneme ”P3” follows phoneme ”P2”. For the purpose, the system again use HLEd tool of HTK to expand Mono-phone level transcription to Tri-phone level transcription.

4.5 Acoustical Analysis

In Acoustical Analysis, the speech files obtained are represented in more compact and efficient way by extracting features of speech files. The system uses the Mel Frequency Cepstral Coefficients (MFCCs) to extract features from speech files. This is done with the help of HCopy tool of HTK. HCopy tool uses a configuration file that is used for setting the parameters of the acoustical coefficient extraction [1]. Parameterization of Speech signals is done by using 39 MFCC features. The 39 features include an energy coefficient, twelve MFCC coefficients and their first and second derivatives. The speech signals are divided into frames having length 25ms and a windowing function is used that is multiplied with every frame.

Table 4.1: Words and their Corresponding Phonemes used in dictionary of the system

Word	Phonemes	Word	Phonemes	Word	Phonemes
cipher	s ih ph a r	painti	p eh n t ih	sattar	s a tt a r
ikk	ih k	chhatti	ch a t ih	ikhttar	ih k a h a tt a r
do	d ow	sainti	s eh n t ih	bahttar	b a h a tt a r
tinn	t ih n	athtti	a th t ih	tihattar	t ih h a tt a r
char	ch ah r	untali	u n t ah l ih	chuhattar	ch u h a tt a r
panj	p ah n j	chali	ch ah l ih	panjhattar	p a ch a tt a r
chhe	ch eh	iktali	ih k t ah l ih	chhihattar	ch ih h a tt a r
satt	s ah tt	batali	b a t ah l ih	satattar	s a t a tt a r
atth	ah th	tartali	t a r t ah l ih	athattar	a th a tt a r
naum	n ow	chutali	ch ow t ah l ih	unasi	u n ah s ih
das	d a s	pantali	p a n t ah l ih	assi	a s ih
giaram	g ih ah r ah	chhiali	ch ih ah l ih	ikasi	ih k ah s ih
baram	b ah r ah	santali	s a n t ah l ih	biasi	b ih ah s ih
teram	t eh r ah	athtali	a td a t ah l ih	tariasi	t a r ih ah s ih
chaudam	ch ow d ah	unanja	u n a n j ah	churasi	ch u r ah s ih
pandram	p a n d r ah	panjah	p a n j ah	pachai	p a ch ah s ih
solam	s ow l ah	ikvanja	ih k v a n j ah	chhiasi	ch ih ah s ih
sataram	s a t ah r ah	bavanja	b a v a n j ah	satasi	s a t ah s ih
atharam	a th ah r ah	tarvanja	t a r v a n j ah	athasi	a th ah ih
unni	u n ih	churanja	ch u r a n j ah	unanve	u n ah n v eh
vih	v ih	pachvanja	p a ch v a n j ah	nabbe	n a b eh
ikki	ih k ih	chhapanja	ch a p a n j ah	ikanvem	ih k ah n v eh
bai	b ah ih	satvanja	s a tt v a n j ah	banvem	b ah n v eh
tei	t eh ih	athvanja	a th a v a n j ah	tirianavan	t ih r ih ah n v eh
chauvi	ch ow v ih	unahat	u n ah a th	churranavan	ch u r ah n v eh
pachi	p a ch ih	satth	s a th	pachannaven	p a ch ah n v eh
chhabbi	ch ah b ih	ikahath	ih k ah a th	chhiannaven	ch ih ah n v eh
satai	s a t ah ih	bahath	b ah a th	satannaven	s a t ah n v eh
athai	a th ah ih	trehat	t r eh ah a th	athannaven	a th ah n v eh
untti	u n t ih	chaunhat	ch ow ah a th	narhinnaven	n a td ih n v eh
tih	t ih	pehant	p eh ah a th	sau	s ow
iktti	ih k t ih	chhehat	ch eh ah a th		
batti	b a t ih	satahat	s a t ah a th		
teti	t eh t ih	atahat	a th ah a th		
chaunti	ch ow n t ih	unattar	u n a tt a r		

4.6 Training

The System is trained for both sets of data used in database preparation.

For the first set of data, the System is trained by using 760 speech files. For the purpose, Macro definition also known as prototype is created by the system for HMM ”proto” in the form of text description file that uses six states for representation of word. HMM “proto” is used only for flat initialization that is done by using HCompV tool of HTK. Flat initialization means every state of HMM has same mean and variance vectors. HCompV tool initialize the HMM “proto” by modifying its values. The modified HMM ”proto” is used by the system to obtain three Master Macro files (.mmf) [1] for word level, mono-phone level and tri-phone level recognition respectively. After creating .mmf files for the three levels of recognition, HERest tool of HTK that works on Baum-Welch Algorithm [32] is used by the system to re-estimate the values of HMMs at the three levels. HERest tool is used 17 times to re-estimate the values of HMMs at every level which means at word level, at mono-phone level and at tri-phone level.

For the second set of data, the system uses the same procedure for training that it uses for training in the first set of data.

CHAPTER-5
TESTING AND EXPERIMENTAL OBSERVATIONS

5.1 Testing

The System is tested for both sets of data used in database preparation.

For the first set of data, the system uses 250 speech files for testing out of which 50 speech files are taken from known speaker as the speaker is also involved in training and 200 speech files are taken from two unknown speakers (100 speech files from both) as they are not involved in training. The system uses HVite tool of HTK for the purpose. The HVite tool uses Viterbi algorithm [2] that is used for matching each input speech file with one of HMM. During testing, the HVite tool store result in a .mlf file. By using HVite tool at the three levels that is word level, mono-phone level and tri-phone level, the system obtain nine .mlf files for the three levels (three .mlf files for every level for the three speakers respectively). The generated .mlf files at word level contains recognized words corresponding to input speech files. Similarly, the generated .mlf files at mono-phone level and tri-phone levels contain recognized monophones and triphones corresponding to input speech files.

For the second set of data, the system uses the same procedure for testing that it uses for testing in the first set of data.

5.1.1 Graphical User Interface for testing the system

A Graphical User Interface (GUI) is developed in Netbeans 8.0 using language Core Java. GUI enables the users to view transliteration of Punjabi word in English corresponding to HTK (.mfcc) file and to view the results of testing at the three levels by the three speakers used for testing for second set of data as its accuracy is higher than that of first set of data.

Graphical User Interface provides three facilities to the users that are explained in following sub-sections.

5.1.1.1 View Transliteration of Punjabi Word in English Corresponding to HTK file

User can view transliteration of Punjabi Word in English Corresponding to HTK (.mfcc) file. Figure 5.1 and Figure 5.2 shows how user can get Punjabi word corresponding to HTK file.

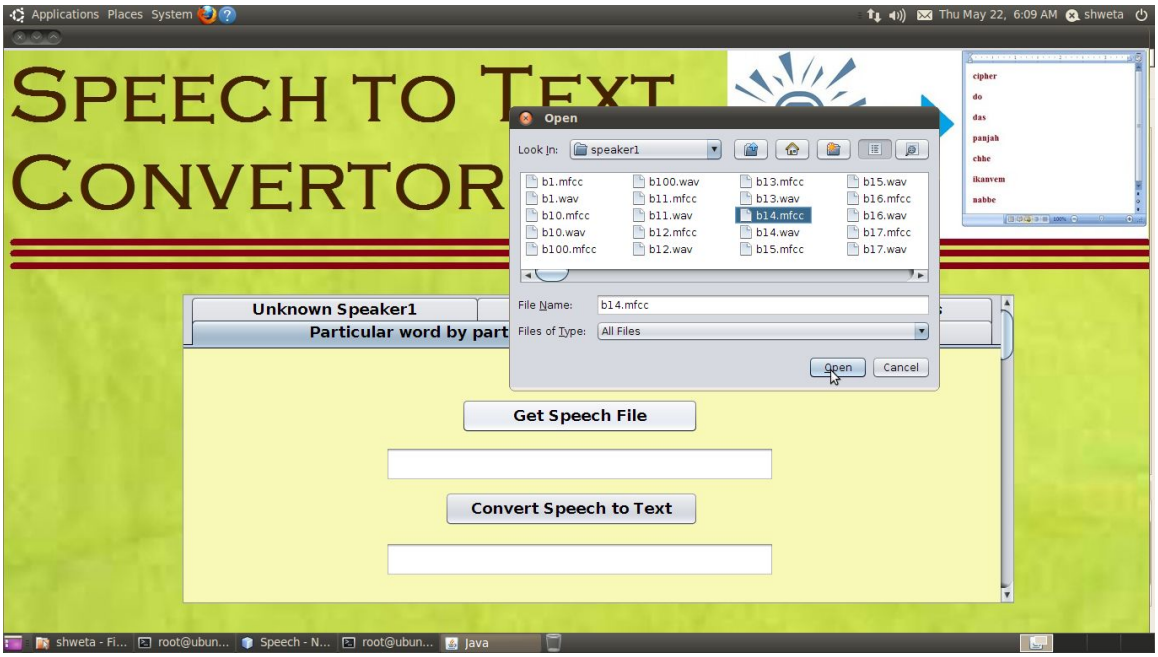


Figure 5.1: Dialog box with the help of which user can open HTK (.mfcc) file

5.1.1.2 View Expected words and their Corresponding Recognized words at the three levels

User can view transliteration of expected Punjabi words in English and their corresponding recognized words at the three levels by the speakers (Known Speaker as the Speaker is involved in training and Unknown Speaker 1 and Unknown Speaker 2 as both are not involved in training) used for testing the system. Figure 5.3, Figure 5.4 and Figure 5.5 shows the expected and their corresponding words by the three speakers respectively.

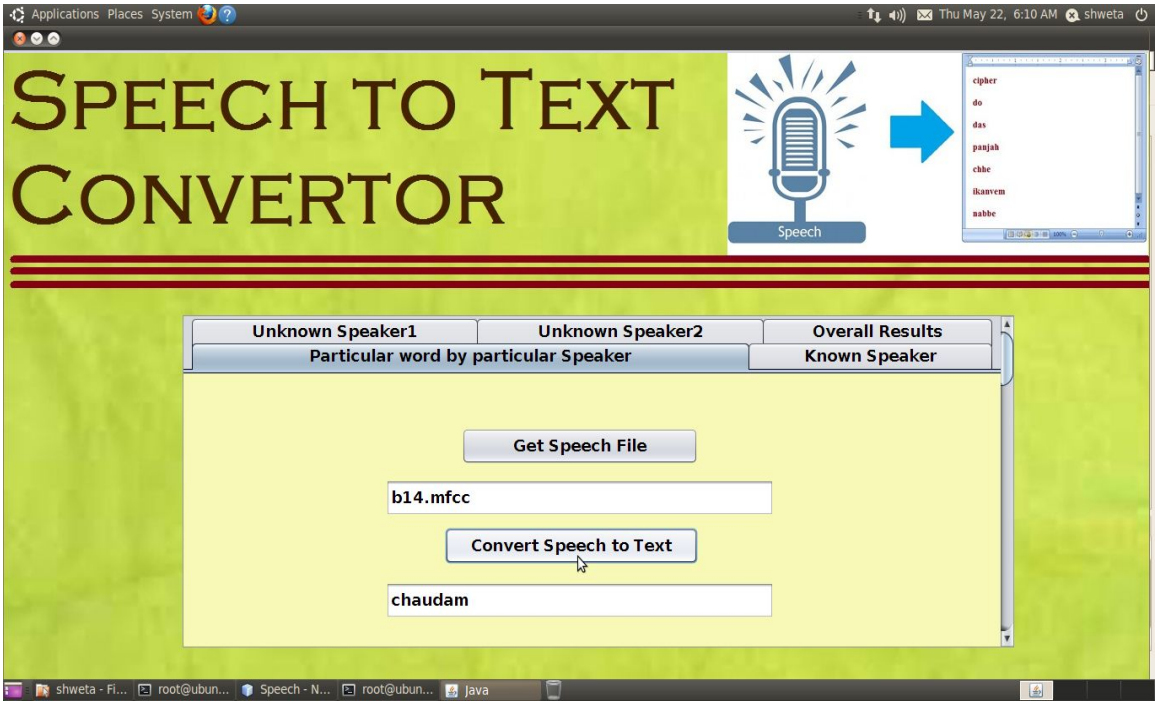


Figure 5.2: Transliteration of Punjabi word corresponding to selected HTK file

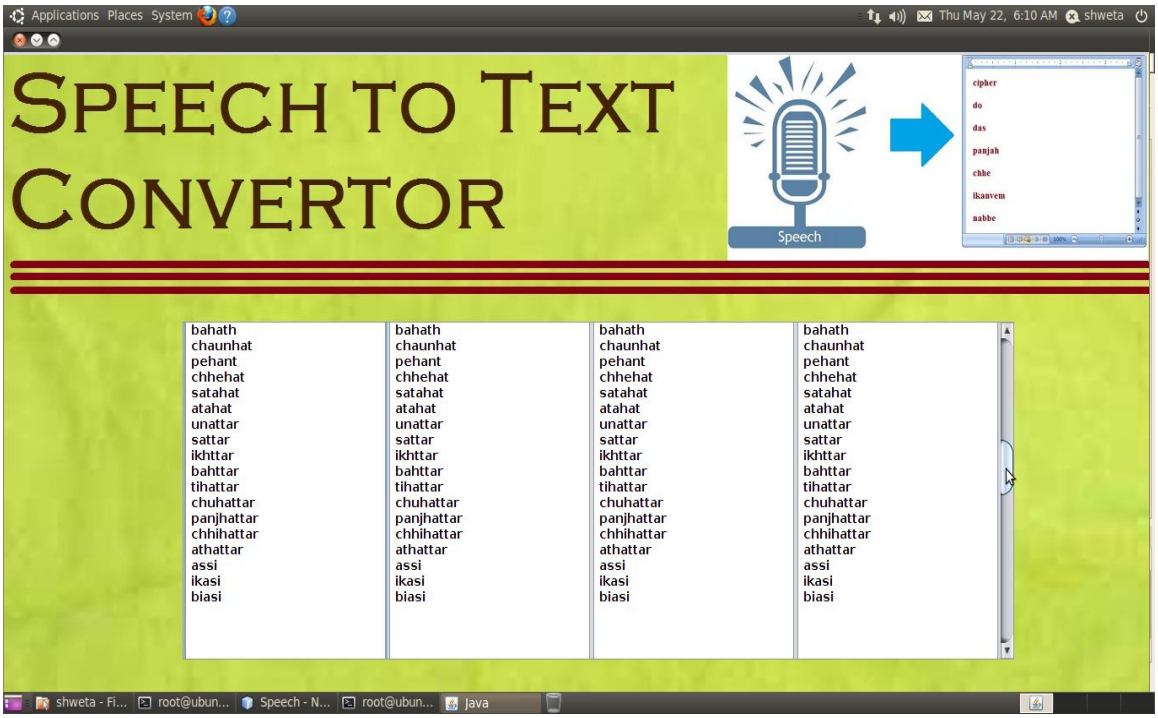


Figure 5.3: Expected and their Corresponding recognized words by Known Speaker

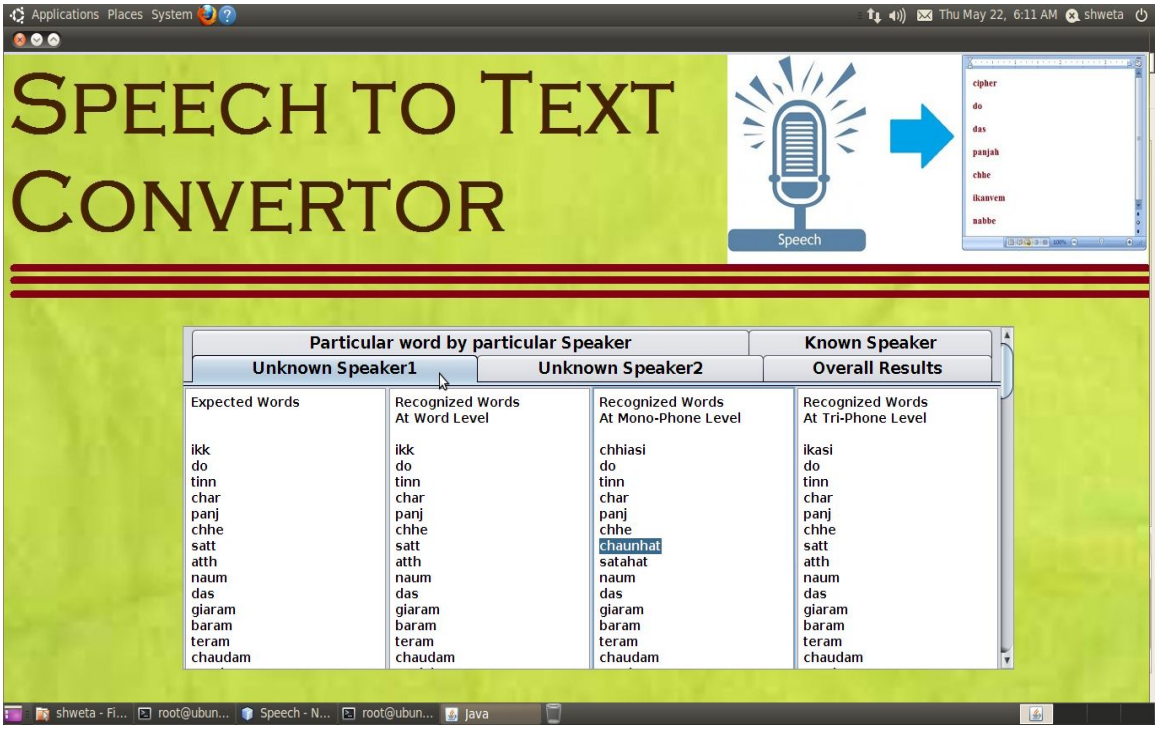


Figure 5.4: Expected and their Corresponding recognized words by Unknown Speaker 1

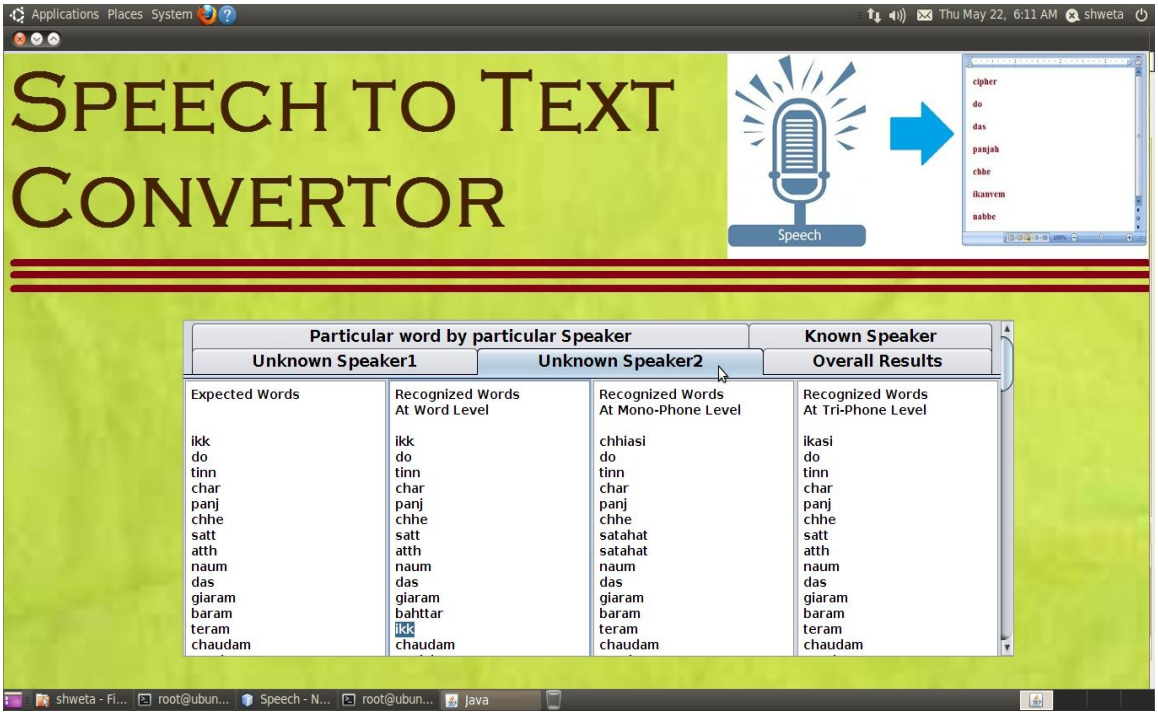
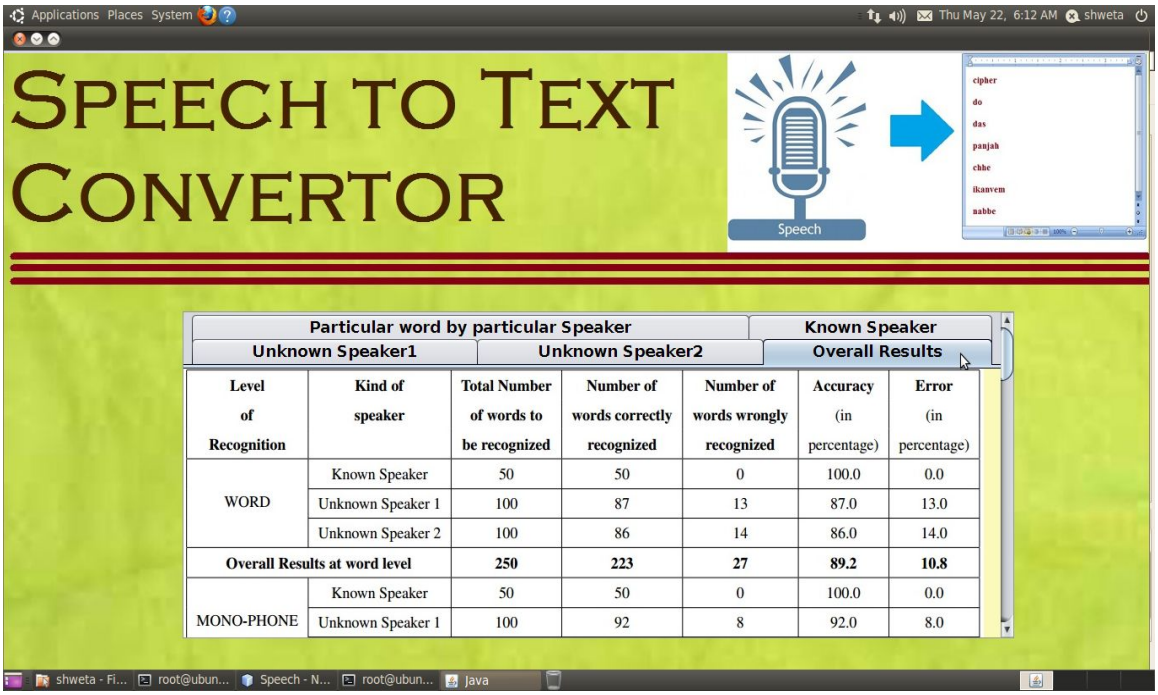


Figure 5.5: Expected and their Corresponding recognized words by Unknown Speaker 2

5.1.1.3 View Accuracy of the system at the three levels

User can view the accuracy of the system at the three levels by the three speakers and overall accuracy of the system at the three levels. Figure 5.6 shows the accuracy of the system at the three levels.



Level of Recognition	Kind of speaker	Total Number of words to be recognized	Number of words correctly recognized	Number of words wrongly recognized	Accuracy (in percentage)	Error (in percentage)
WORD	Known Speaker	50	50	0	100.0	0.0
	Unknown Speaker 1	100	87	13	87.0	13.0
	Unknown Speaker 2	100	86	14	86.0	14.0
Overall Results at word level		**250**	**223**	**27**	**89.2**	**10.8**
MONO-PHONE	Known Speaker	50	50	0	100.0	0.0
	Unknown Speaker 1	100	92	8	92.0	8.0

Figure 5.6: Accuracy of the system at the three levels

5.2 Experimental Observations

The system uses HResults tool of HTK to get the accuracy of the system for both sets of data. For both sets of data, HResults tool compare the nine .mlf files generated while testing at the three levels with their transcriptions respectively to check whether words, monophones and triphones recognized in .mlf files during testing are recognized correctly or not. Table 5.1 shows the results obtained by HResults tool for the three levels that is word level, mono-phone level and tri-phone level respectively for first set of data. Table 5.2 shows the results obtained by HResults tool for the three levels that is word level, mono-phone level and tri-phone level respectively for second set of data. Figure 5.7 shows the accuracy of the system at the three levels for the three speakers used for testing for both sets of data. Figure 5.8 shows the overall accuracy of the system at the three levels for both sets of data.

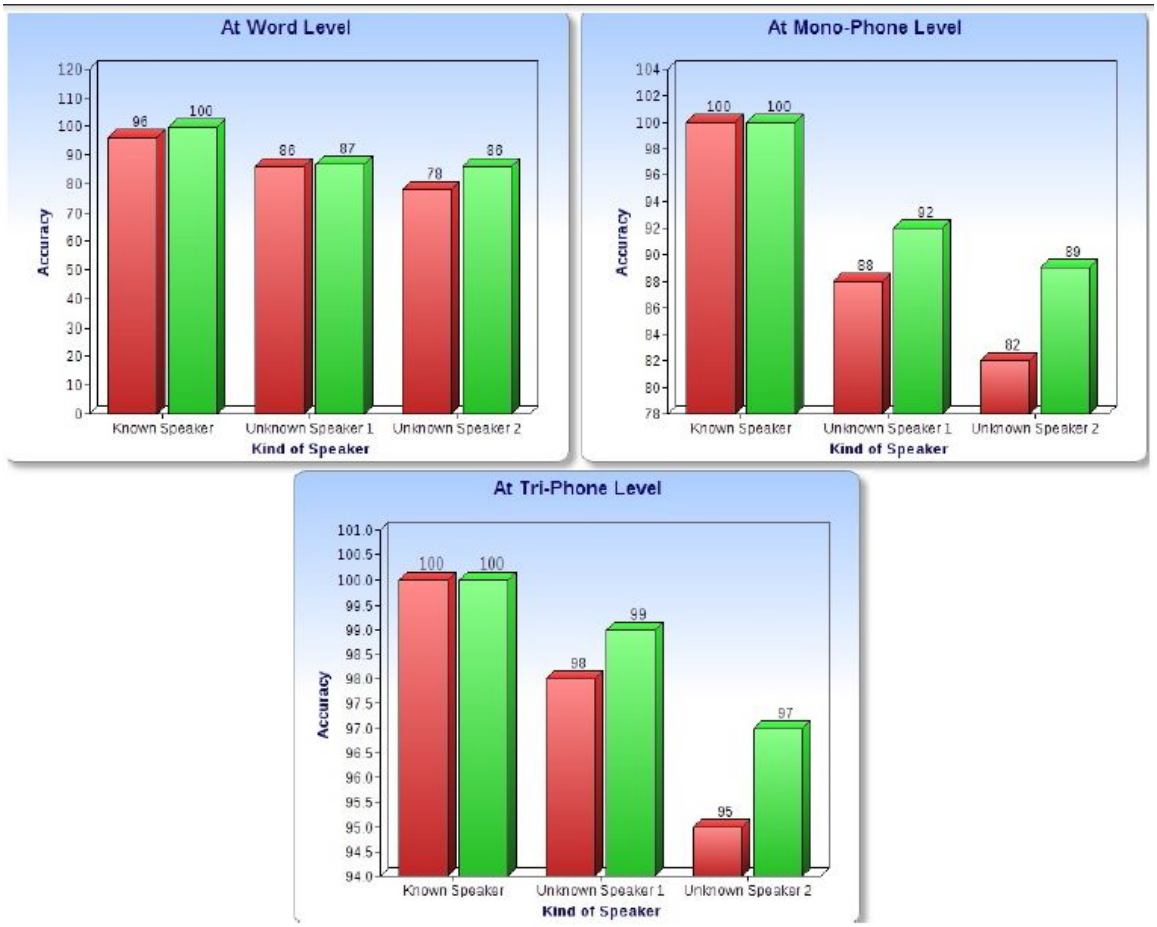


Figure 5.7: Graphs showing accuracy of the system at Word level, Mono-phone level and Tri-phone level for the three speakers used for testing

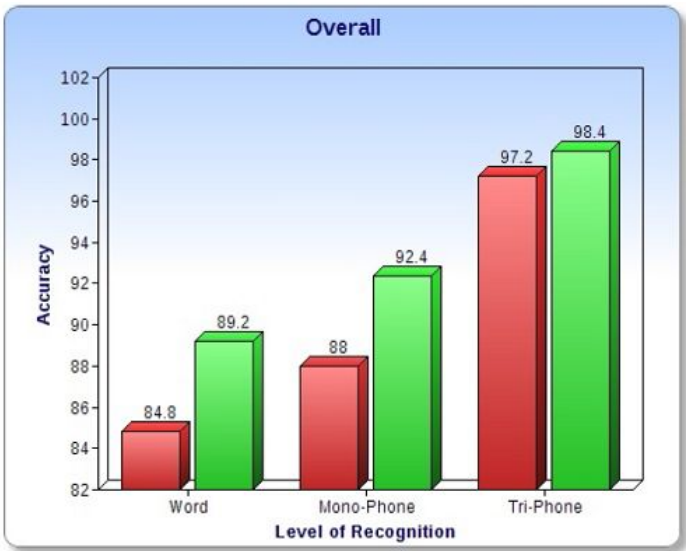


Figure 5.8: Graph showing accuracy of the system at the three levels of recognition

Table 5.1: Experimental Observations for first set of data

Level of Recognition	Kind of speaker	Total Number of words to be recognized	Number of words correctly recognized	Number of words wrongly recognized	Accuracy (in percentage)	Error (in percentage)
WORD	Known Speaker	50	48	2	96.0	4.0
	Unknown Speaker 1	100	86	14	86.0	14.0
	Unknown Speaker 2	100	78	22	78.0	22.0
Overall Results at word level		250	212	38	84.8	15.2
MONO-PHONE	Known Speaker	50	50	0	100.0	0.0
	Unknown Speaker 1	100	88	12	88.0	12.0
	Unknown Speaker 2	100	82	18	82.0	18.0
Overall Results at mono-phone level		250	220	30	88.0	12.0
TRI-PHONE	Known Speaker	50	50	0	100.0	0.0
	Unknown Speaker 1	100	98	2	98.0	2.0
	Unknown Speaker 2	100	95	5	95.0	5.0
Overall Results at tri-phone level		250	243	7	97.2	2.8

Table 5.2: Experimental Observations for second set of data

Level of Recognition	Kind of speaker	Total Number of words to be recognized	Number of words correctly recognized	Number of words wrongly recognized	Accuracy (in percentage)	Error (in percentage)
WORD	Known Speaker	50	50	0	100.0	0.0
	Unknown Speaker 1	100	87	13	87.0	13.0
	Unknown Speaker 2	100	86	14	86.0	14.0
Overall Results at word level		250	223	27	89.2	10.8
MONO-PHONE	Known Speaker	50	50	0	100.0	0.0
	Unknown Speaker 1	100	92	8	92.0	8.0
	Unknown Speaker 2	100	89	11	89.0	11.0
Overall Results at mono-phone level		250	231	19	92.4	7.6
TRI-PHONE	Known Speaker	50	50	0	100.0	0.0
	Unknown Speaker 1	100	99	1	99.0	1.0
	Unknown Speaker 2	100	97	3	97.0	3.0
Overall Results at tri-phone level		250	246	4	98.4	1.6

CHAPTER-6
CONCLUSION AND FUTURE SCOPE

CONCLUSION

In Conclusion,

- Noise Reduction Technique (Auto Spectral Subtraction) has been applied that improves the accuracy of the system to some extent.
- Below two points clearly give the comparison of the system at two different aspects. One when no noise reduction technique is applied and other when noise reduction technique is applied.
- HMM Based Speaker Independent Isolated Word Speech to Text Conversion System is developed with accuracy of 84.8% at word level, 88% at mono-phone level and 97.2% at tri-phone level for the set of data in which noise reduction is not applied to the speech files
- HMM Based Speaker Independent Isolated Word Speech to Text Conversion System is developed with accuracy of 89.2% at word level, 92.4% at mono-phone level and 98.4% at tri-phone level for the set of data in which **noise reduction technique (Auto Spectral Subtraction)** is applied.
- The accuracy obtained in the system is still not 100% accurate. Disturbances caused by environment and other causes, different styles of speaking used by humans are the major reasons that existing speech recognition systems are not 100% accurate.

FUTURE SCOPE

To further improve the accuracy of the system, following steps can be done.

- Different noise compensation/speech enhancements techniques can be used to make the system more accurate.
- Accuracy of the system also depends on training process of the system. For the purpose, HTK tool HERest can be modified.

The System is limited to recognition of isolated words and vocabulary of 1010 words, following steps can be done.

- The system can be further extended to connected word recognition or continuous word recognition or spontaneous word recognition.
- The system can be further extended to very large vocabulary size.

CHAPTER-7
REFERENCES

[1] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu , et al., "A tutorial example of using htk," in *The HTK Book*. Cambridge University Engineering Department, 2002, pp. 23–66.

[2] D. Jurafsky and J. H. Martin, "HMMs and Speech Recognition," in *Speech and Language Processing,* S. Russell and P. Norving, Dorling Kindersley Pvt. Ltd., India, 2000, pp. 261-309.

[3] J. Beh and K. Hanseok, "Spectral Subtraction Using Spectral Harmonics for Robust Speech Recognition in Car Environments", *Springer-Verlag Berlin Heidelberg*, 2003, pp. 1109-1116.

[4] B. A. Forouzan, "Data and Signals," in *Data Communications and Networking*, McGraw-Hill Companies, 4th edition, 2007, pp. 57-88.

[5] S. J. Young, P. C. Woodland and W. J. Byrne , "Spontaneous speech recognition for the credit card corpus using the htk toolkit," *IEEE transaction on Speech and Audio Processing*, vol. 2, October 1994, pp. 615–621.

[6] T. Hain, P. C. Woodland, G. Evermann, D. Povery, "New features in the cu-htk system for transcription of conversational telephone speech," Cambridge University Engineering Department, Cambridge, CB2 1PZ, UK, 2001.

[7] O. Majdalawieh, J. Gu and M. Meng, "An htk-developed hidden markov model (hmm) for a voice-controlled robotic system," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, October 2004, pp. 4050–4055.

[8] R. Sinha, M. J. F Gales, X. A. Liu, K. C. Sim and P.C. Woodland, "The cu-htk mandarin broadcast news transcription system," *International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. 1077–1080.

[9] S. Manadhar, B. Zi'olko, R. C. Wilson, M. Zi'olko and J. Galka, "Application of htk to the polish language," *International Conference on Audio, Language and Image Processing*, 2008, pp. 1759–1764.

[10] C. Zhong and Z. Miao, "Pronunciation recognition and assessment for mandarin chinese," *Congress on Image and Signal Processing*, 2008, pp. 352–356

[11] S. Mandal, B. Das and P. Mitra, "Bengali speech corpus for continuous automatic speech recognition system," Indian Institute of Technology computer Science and Engineering Department, Kharabpur, India, 2011.

[12] Z. Du[1], X. Li[2] and J. W[3], "Accelerating the Training of HTK on GPU with CUDA," [1]Tsinhua National Laboratory for information Science and Technology, Department of Computer Science and Technology, Tsinghua University 100084, [2]School of Computer, Beijing University of Posts and Telecommunications, 100876, [3]Tsinhua National Laboratory for information Science and Technology Department of Electronic Engineering, Tsinghua University, 100084, Beijing, China, 2012.

[13] W. Ghai and N. Singh, "Phone based acoustic modeling for automatic speech recognition for punjabi language", *Journal of Speech Sciences*, 2013, pp. 69–83

[14] S. Sultana et al., "Bangla Speech-to-Text Conversion using SAPI", *International Conference on Computer and Communication Engineering (ICCCE 2012)*, Kuala Lumpur, Malaysia, 3-5 July 2012, pp. 385–390.

[15] P. Heracleous et al., "Visual-speech to text conversion applicable to telephone communication for deaf individuals", *18th International Conference on Telecommunications*, 2011, pp. 130–133.

[16] P. C. Woodland *et al.*, "Broadcast News Transcription using HTK," Cambridge University Engineering Department, Cambridge, CB3 1PZ, England, 1997.

[17] W. Zhou *et al.*, "A Comparison between HTK and SPHINX on Chinese Mandarin," *International Joint Conference on Artificial Intelligence,* Department of Electrical Engineering, Beijing Normal University, Beijing, China, 2009.

[18] B. A. Qatab and R. N. Ainon, "Arabic Speech Recognition using Hidden Markov Model Toolkit (HTK)," Software Engineering Department, University Of Malaya Kuala Lumpur, Malaysia, 2010.

[19] N. H. Quang et al., "Automatic Speech Recognition for Vietnamese using HTK system," School of Information and Communication Technology, Hanoi University of Technology, Hanoi, VIETNAM, 2010.

[20] M. Dua[1], R. K. Aggarwal[2], V. Kadyan[3] and S. Dua[4], "Punjabi automatic speech recognition using HTK," [1,2]Department of Computer Engineering, NIT, Kurukshetra, India, [3]Department of Computer Engineering, DIET, Karnal, India, [4]Department of Electronics and Communication Engineering, RPIIT, Karnal, India, July 2012.

[21] S. C. Pammi, V. Keri, "HTKTrain: A Package for Automatic Segmentation," Computer Science and Engineering Department, IIIT, Hyderabad.

[22] R. Das and P. K. Das, "Design and Implementation of Monophones and Triphones-Based Speech Recognition Systems for Voice Activated Telephony," BIJIT - BVICAM's International Journal of Information Technology Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi, India, June 2013.

[23] M. Katz, A. Mbogho, "Speech Recognition Across South African Accents," Computer Science Department, University of Cape Town, 2009.

[24] A. Pittermann and J. Pittermann, “Getting Bored with HTK? Using HMMs for Emotion Recognition from Speech Signals,” Department of Information Technology, University of Ulm, 89069 Ulm, Germany, 2006

[25] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang and P. C. Woodland, “Development of the 2003 CU-HTK Conversational Telephone Speech Transcription System,” Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK, 2004

[26] D. Y. Kim, H. Y. Chan, G. Evermann, M. J. F. Gales, D. Mrva, K. C. Sim and P. C. Woodland, “Development of the CU-HTK 2004 Broadcast News Transcription Systems,” Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ, UK, 2005.

[27] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, “Progress in the CU-HTK Broadcast News Transcription System,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, 5 September, 2006, pp. 1513-1525.

[28] G. Zhang, J. Yin, Q. Liu and C. Yang, “The Fixed-Point Optimization of Mel Frequency Cepstrum Coefficients for Speech Recognition,” School of Applied Sciences, Harbin University of Science and Technology, Harbin, China, 2011.

[29] P. V. Bhaskar, S. R. M. Rao and A.Gopi, “HTK Based Telugu Speech Recognition,” *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, Issue 12, India, December 2012.

[30] J. Kaur, R. Kaur, et al., “Issues Involved In Speech To Text Conversion,” *International Journal of Computational Engineering*, ISSN: 2250–3005, Vol. 2, Issue No. 2, Mar-Apr 2012, pp. 512-515.

[31] P. Saini, P. Kaur and M. Dua, "Hindi Automatic Speech Recognition Using HTK," *International Journal of Engineering Trends and Technology (IJETT)*, Vol. 4, Issue 6, India, June 2013

[32] L. Moss. A tutorial on Example of the Baum-Welch Algorithm, Q520, Spring 2008.

CHAPTER-8
LIST OF PUBLICATIONS

Accepted Paper

1. Shweta Mittal and Karun Verma, "HMM based Punjabi Speaker Independent Isolated Word Speech to Text Conversion using HTK," accepted in *Second International Conference on Emerging Research in Computing, Information, Communication and Applications, ERCICA 2014* (Sponsored by Elsevier), Nitte Meenakhsi Institute of Technology, Bangalore, India, 01-02 August, 2014.