Automation of Pitch Accent Markings in Punjabi Speech

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

Master of Technology

in

Computer Science and Applications

Submitted By

Minakshi Bansal

(Roll No. 601303018)

Under the supervision of:

Dr. R.K. Sharma

Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Automation of Pitch Accent Markings in Punjabi Speech*", in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Applications* 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 *Dr. R.K. Sharma* 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.

Minakshi Bansal

(Minakshi Bansal)

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

(R.K. Sharma)

Professor, CSED

Countersigned by

(Dr. Deepak Garg)

Head

Computer Science and Engineering Department

Thapar University

Patiala

(Dr. S. S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

ACKNOWLEDGEMENTS

A dissertation cannot be completed without the people around me who helped me directly or indirectly. They made this period a real educative, pleasurable and memorable one. My sincere thanks to all those people.

Firstly, I would like to thank my supervisor **Dr. R.K. Sharma**, for giving me an opportunity to work under his guidance. His constant support, guidance and efforts helped me a lot in completing this research work. His experience gave me a better incite to work.

I would also like to offer my sincere thanks to all faculty members, teaching and non-teaching staff of Thapar University for their assistance.

I would also like to thank my parents and friends for their encouragement during the entire course of this work.

Above all, I owe my reverence to Almighty for the kindness who blessed me at finish of whole work.

Minakshi Bansal

(Minakshi Bansal)

ii

ABSTRACT

Speech has been one of the major forms of human communication and a unique characteristic of the human species. Speech is produced by vibrations of the vocal cords. The use of prosodic knowledge in automatic speech recognition is a widely researched topic in recent years. To incorporate prosody knowledge into speech recognition by marking the variations in pitch is the main objective of this thesis. The rate of vibration of vocal cords is called fundamental frequency or pitch. Pitch marking is related to the instances of glottal closure. An accurate pitch marking directly influences the quality of speech signal. For pitch marking, voiced and unvoiced regions are separated because pitch marking is applicable only to the voiced parts of speech, no pitch can be detected in case of unvoiced speech. A GUI has also been developed for automatic prosody labeling of pitch with variation of pitch labels. This thesis is divided into five chapters. A brief outline of these chapters is given below.

Chapter 1 includes the basic terminology and tools which one needs to be familiar with for prosody labeling of pitch.

Chapter 2 includes literature survey. This chapter is divided into three parts. First part deals with the review of literature of basic fundamentals of speech. Second part deals with the review of literature of prosody labeling systems and the last part deals with the review of literature of pitch marking techniques in voiced speech segments.

Chapter 3 discusses the problem statement and also includes the data collection for pitch marking. Total duration of data considered in this work is 32 minute and 13 seconds. A total of 40 files are taken from 10 different speakers. Each speaker has contributed 4 files.

Chapter 4 focuses on the algorithm used for pitch marking. This chapter also includes the outputs of the algorithm and results are presented in different figures. It shows the output of algorithm on 40 files, manual pitch marking on 40 files by 4 users and the comparison between them.

Chapter 5 presents the conclusion of the work done and the further scope of this work for increasing the accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Expanded Form
F	Flat region
H	High region
L	Low region
VH	Very high region
VL	Very low region
LH	Low to high region
HL	High to low region
VLH	Very low to high region
HVL	High to very low region
VHL	Very high to low region
LVH	Low to very high region
HE	Human Evaluator
SVM	Support Vector Machine
GSST	German Spontaneous Scheduling Task
GUI	Graphical User Interface
TTS	Text To Speech
CART	Classification and Regression Tree

TIMIT	Texas Instruments and Massachusetts Institute of Technology
NTIMIT	Network TIMIT
TD-PSOLA	Time Domain Pitch Synchronous Overlap and Add
RDP	Restricted Dynamic Programming
ACF	Auto Correlation Function
HTK	Hidden Markov Model Toolkit
TVs	Vocal Tract Constriction Variables

CHAPTER 1

Introduction

Speech is considered as one of the major forms of human communication and a unique characteristic of the human species. Due to speech recognition, computers can follow human voice commands and perform accordingly. Long time back, technical understanding of the processes involved in it was not there but due to the advances in technology, researchers have developed a great amount of understanding in speech. The use of prosodic knowledge in automatic speech recognition is an attractive topic of research in recent years.

One of the major advances in recent years is understanding the speech production. When inhaled air from lungs is pushed, vibration of vocal cords occurs and speech sound is produced. The major speech production organs are – lungs, trachea and glottis. Voiced speech is produced by continuous vibration of vocal cords. Firstly, the vocal cords come closer, which temporarily blocks the flow of air from lungs and leads to increased glottal pressure. When this glottal pressure becomes greater than the resistance offered by the vocal cords, they open again. The folds open and close rapidly due to the number of factors, such as elasticity, the Bernoulli effect, etc. When they open and close again and again, air flow through the glottal opening. The fundamental frequency is determined by the frequency of these pulses and thus, contributes to the perceived pitch of the produced sound (Dikshit, 2000).

The main objective of this thesis is to incorporate prosody knowledge into speech recognition by marking the variations in pitch. Pitch accent marking is one of the important parameters or feature of sound signals. It can be done for the voiced parts of a speech. For pitch marking, the sampling rate of the signal should not be more than 8000 Hz. If it is so, it is resampled to 8000 Hz (Audacity 2.1.1 Manual, 2015).

Pitch marking is an important task for prosody modifications in speech synthesis. It is related to the instances of glottal closure. An accurate pitch marking directly influences the quality of speech signal (Sreejith *et al.,* 2013). Pitch marking is done through seven parameters in this work. The symbol F has been used for flat region of the speech signal. The symbol LH has been used for low to high region of the speech signal. The symbol HL has been used for high to low region of the speech signal. The symbol VLH has been used for very low to high region of the speech signal. The symbol HVL has been used for high to very low region of the speech signal. The symbol VHL has been used for very high to low region of the speech signal. The symbol LVH has been used for low to very high region of the speech signal.

A GUI has also been developed for this purpose.

1.1 Categories of Speech

1.1.1 Voiced Region

During production of voiced speech, air which is exhaled out of lungs through trachea is interrupted periodically by vibrating vocal folds i.e. the vocal folds vibrate periodically in case of voiced speech. Due to this, a glottal wave is generated that excites the speech production system resulting in voiced speech. Vowels which are an important part of any speech and parts of many consonants are included in voiced speech. Pitch is an important feature of voiced speech (Sharma, 2012).

If the speech signal waveform looks nearly periodic in nature, then it can be considered as voiced speech.

1.1.2 Unvoiced Region

During production of unvoiced speech, air which is exhaled out of lungs through trachea is not interrupted by vibrating vocal folds. Starting from glottis, somewhere along length of vocal tract, total or partial closure occurs which results in obstructing air flow completely or narrowly. As a result, vocal tract system is excited to produce unvoiced speech. It includes many consonants such as “s”, “sh”, etc. In unvoiced speech, usually, no pitch can be detected (Sharma, 2012).

Unlike voiced speech, unvoiced speech does not have any periodic nature. The resulting speech is random in nature. This is the basic difference between voiced speech and unvoiced speech.

1.1.3 Silence Region

In speech production, voiced and unvoiced speech are produced one after the other separated by a region known as silence region. In this state of speech, there is no excitation and therefore, no sound is being produced. Silence region is having the least energy of all. But silence is the must part of any speech signal. If there is no silence region between voiced and unvoiced speech, speech signal can not be understood and it is very difficult to perceive information from the signal. The duration of silence region is an important factor to be considered for an intelligible speech (Sharma, 2012).
If the energy and the amplitude of signal is very low, then it can be considered as silence.

Figure 1 represents the difference between these three regions by showing the word "shalgam" as a speech signal.

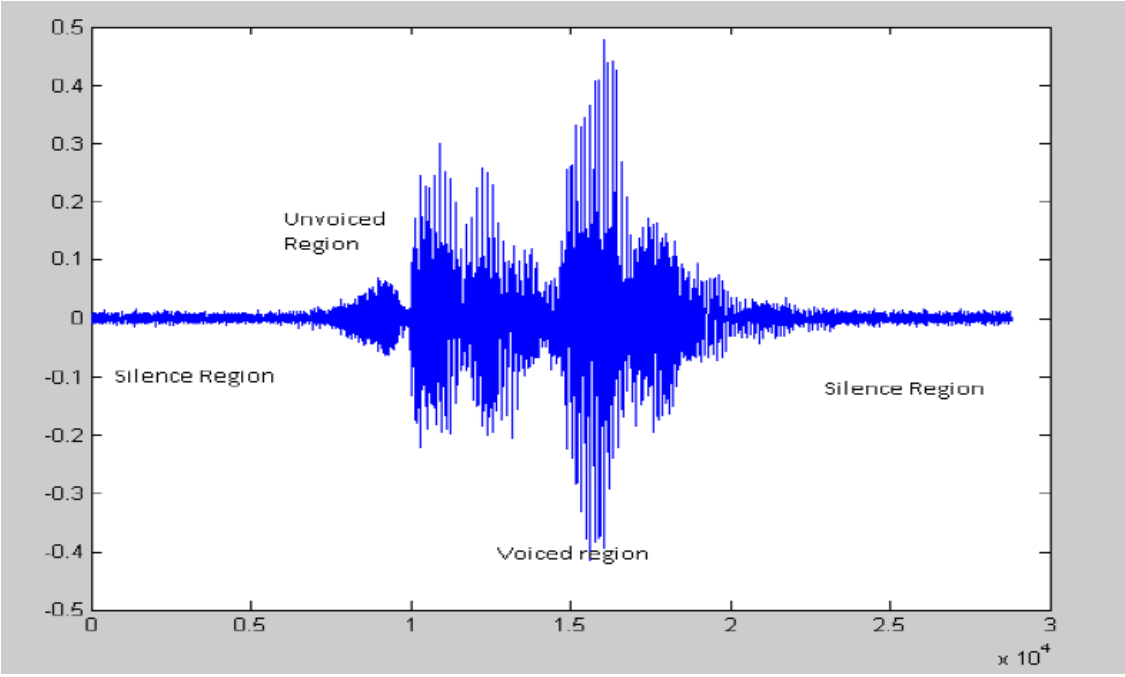

Figure 1: Speech signal waveform of word "shalgam" showing its different regions (Sharma, 2012)

1.2 WaveSurfer Tool

WaveSurfer is an acoustic analysis software package distributed for free by the Centre for Speech Technology (CTT) at the Royal Institute of Technology (KTH) in Stockholm, Sweden. It is very easy to use. WaveSurfer has a simple but powerful interface to work with sound. It can run on most platforms including Microsoft Windows, Mac OS X, Linux, Solaris, HP-UX, FreeBSD and IRIX. It can read and write a number of transcription file formats used in industrial speech research (Hayes, 2010).

When we first open the WaveSurfer, it contains no sound. We can load a sound file from disk or start recording, using the tape-recorder like controls in the upper right corner.

Figure 2 represents the initial interface of WaveSurfer.

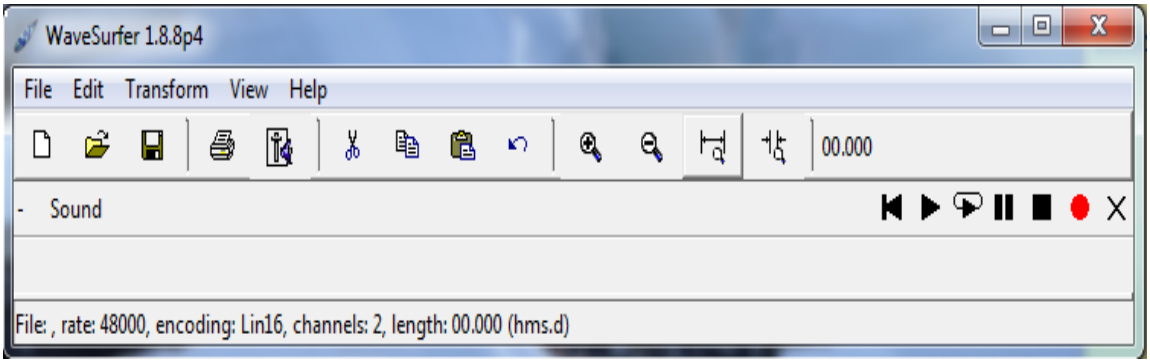

Figure 2: Initial interface of WaveSurfer

Figure 3 shows opening a file in WaveSurfer.

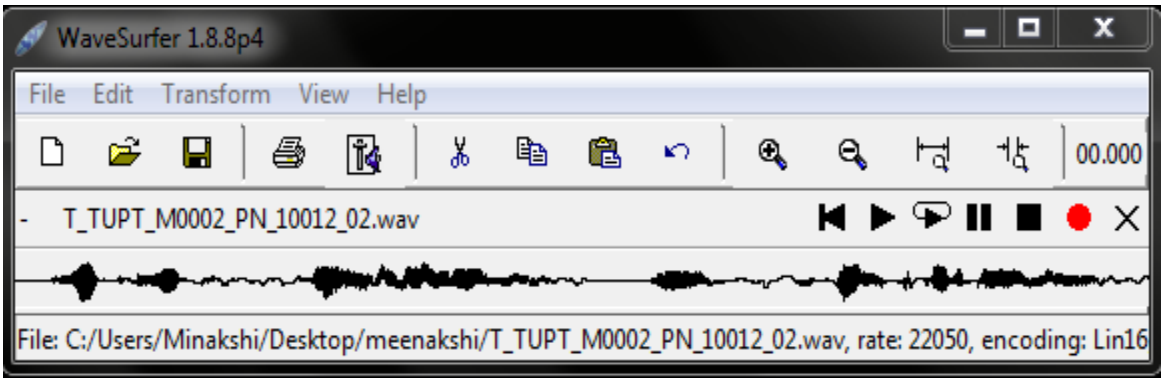

Figure 3: Opening a file

There are various configurations in which the audio file can be opened e.g. ProPEn, Spectrogram, Transcription, etc. So when we open any file in WaveSurfer, it asks for the configuration in which the file is to be opened.

Figure 4 shows different configurations in which the file can be opened in WaveSurfer. The file is to be opened in ProPEn mode.

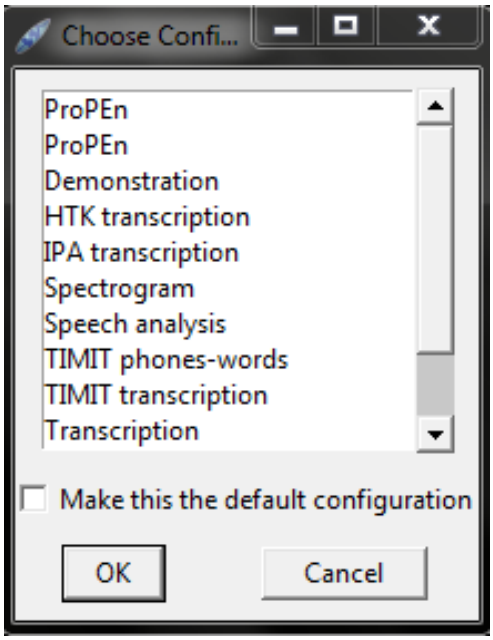

Figure 4: Different configurations

Then, the file gets successfully opened in WaveSurfer as shown in Figure 5.

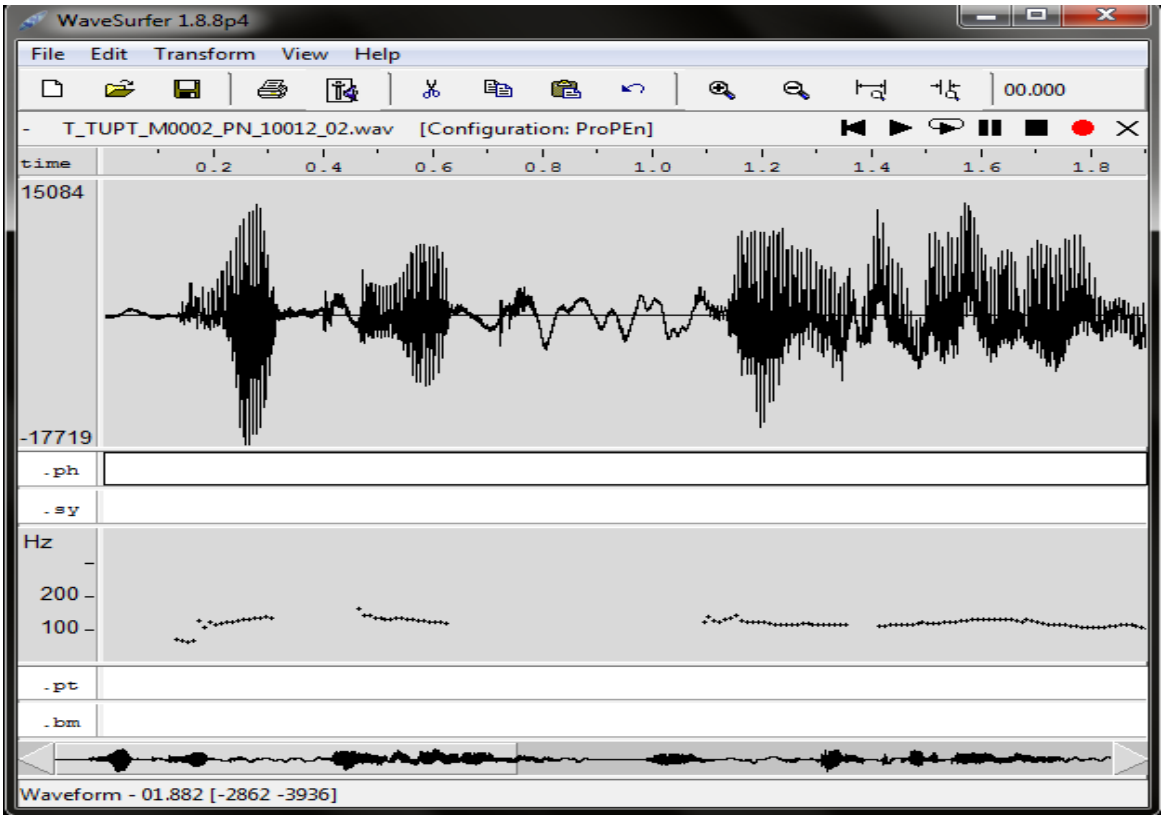

Figure 5: Successfully opened file in WaveSurfer

1.3 Prosody Labeling

Prosody labeling is an important content in linguistic research and voice projects. It is an important element of language. Prosody labeling is done to describe the prosodic events. Incorporating prosodic knowledge into automatic speech recognition improves the performance of the phonetic engine. So, prosody labeling is done which includes marking of syllable boundaries as syllabification, pitch marking and break index marking. Using WaveSurfer software for inserting labels, separate files get created for them.

1.3.1 Syllabification

Syllabification is the process of separation of a word into syllables. Syllable is basically a larger unit as compared to a phoneme. Syllable has a nucleus, normally a vowel sound with optional final and initial margins, generally consonants. Syllabification extension is .sy (Sreejith *et al.*, 2013).

1.3.2 Break Index Marking

Break Index Marking marks the perceived degree of separation between words in the utterance. It marks the duration of silence region in the signal. Four labels are used for break marking - B0, B1, B2, B3 where B0 represents syllable boundary marking with no physical break in the speech signal and B3 corresponds to a full pause such as sentence break.

- B0 – no separation (weakest form of break)
- B1 – word boundaries
- B2 – short pause
- B3 – a full pause, such as at a sentence boundary

Break Marking extension is .bm.

1.3.3 Pitch Marking

Knowledge about the exact moments of excitation of vocal tract is extremely important in speech synthesis and recognition (Sreejith *et al.*, 2013). Pitch accent marking is one of the important features of sound signals. It is applicable only to the voiced parts of speech as no pitch can be detected in case of unvoiced speech. In pitch marking process, one marks epochs (instance of glottal closure) and provide related parameters as fundamental frequency. It determines the exact moments of glottal closure. So, incorporating prosody knowledge into automatic speech recognition through pitch marking is the main objective of this thesis. Pitch variations are labeled automatically to represent the prosody. Variation of pitch provides some melodic properties to speech. For pitch marking, sampling rate of the signal should not be more than 8000 Hz. If so, it is re-sampled to 8000 Hz. .pt is the extension for pitch marking.

Figure 6 shows the pitch marking in WaveSurfer for a wave file of 4 seconds duration.

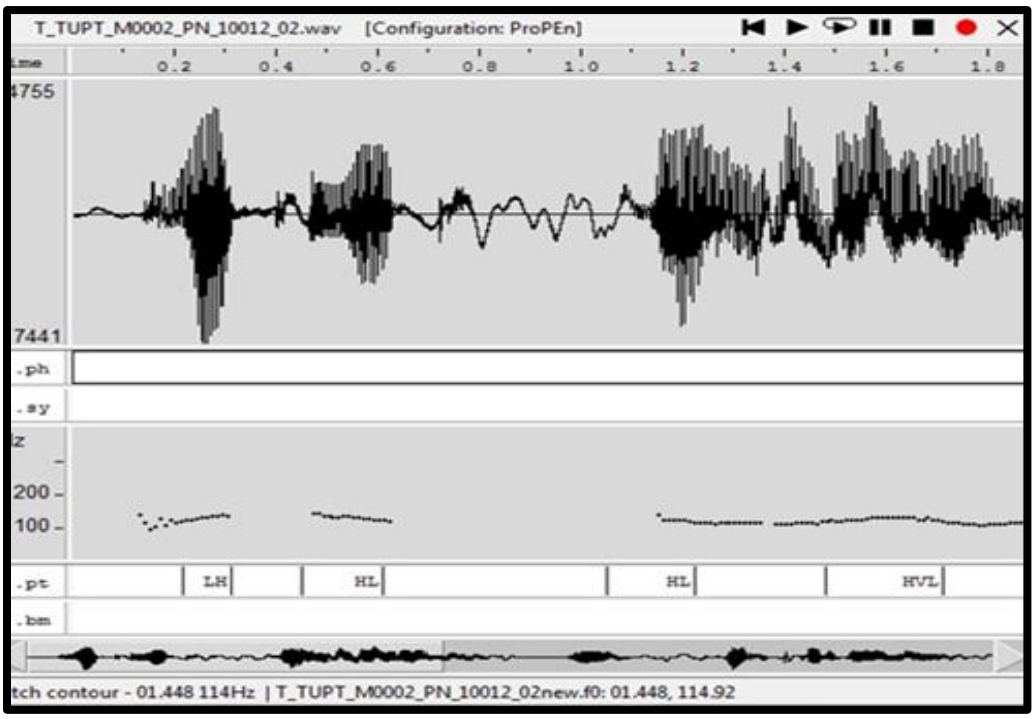

Figure 6: Pitch marking in WaveSurfer for a wave file of 4 seconds duration

1.4 Pitch vs. Tone

Pitch and tone, these two terms are often incorrectly used interchangeably so understanding the difference between them becomes important.

Pitch is an important feature of voiced speech. It represents a particular frequency of sound e.g. 440 Hz. It is a property that allows ordering of sounds on frequency-related scale. It is the rise and fall of our voice. Pitch is basically the quality that allows us to differentiate a sound as relatively 'high' or 'low'. It is determined by the frequency of sound wave vibrations. A higher frequency results in higher pitch and a lower pitch results from a lower frequency.

Unlike pitch which shows the rise and fall of our voice, tone represents the quality of sound, that which differentiates it and makes it recognizable by its constant 'pitch'. Tone deals more with expression in our voice, our emotions and our attitude while delivering

the speech. Tone represents various feelings of anger, sympathy, happiness, sadness etc. when we deliver the speech. A tone's pitch defines its depth (or height) in relation to the complete series of tones that can be heard by the ear. That is why even if the pitch of two instruments is the same, for example, a violin and a flute, they will sound differently. So a lot of factors influence tone, for instance a person's physical condition, breath support, technique and many more (Kivumbi, 2010).

The work presented in this thesis is related to the automation of pitch marking in Punjabi speech.

CHAPTER 2

Literature Survey

In this thesis, literature survey is divided into three parts. First part deals with the review of literature of basic fundamentals of speech. Second part deals with the review of literature of prosody labeling systems and the last part deals with review of literature of pitch marking techniques in voiced speech segments.

2.1 Literature Survey on Speech Fundamentals

Atal *et al.* (1976) described a pattern recognition approach for deciding whether a particular speech segment must be recognized as voiced speech, unvoiced speech, or silence, based on the measurements made on the signal. Five different measurements were made on the speech segment to be classified. The measured parameters were the zero-crossing rate, the speech energy, the correlation between adjacent speech samples, the first predictor coefficient from a 12-pole linear predictive coding analysis, and the energy in the prediction error. The speech segment was assigned to a particular class according to a minimum distance rule. The means and covariances for the Gaussian distribution were determined from manually classified speech data which was included in a training set. It was found that this method provided reliable classification with speech segments as short as 10 ms and gave satisfactory results for both speech analysis-synthesis and recognition applications. Finally, a smoothing algorithm was described where errors in the analysis were corrected, and unusually short intervals were eliminated.

Rabiner *et al.* (1989) proposed connected digit recognition system based on HMMs. Proposed system was trained and tested in three modes: speaker trained, multi-speaker and speaker independent. Evaluation was done on three databases: database widely distributed through National Bureau of Standards, 225 adult talker and 50 talker

connected digit database. 0.78, 2.85 and 2.94 were the string error rate that was observed for all 3 modes.

Brugnara *et al.* (1993) proposed automation of segmentation and labeling of speech of Italian language using HMMs. Training and performance evaluation both were done on TIMIT database. For training purpose 64 speakers were selected, eight sentences from each speaker. Performance was evaluated on 24 different speakers, each speaker was asked to utter eight sentences. It was observed that manual segmentation done by expertise in phonetics provided 93.5% accuracy for locating correct positioned boundary which was not so far from 86.9% obtained by automatic segmentation system.

Slobada *et al.* (1996) proposed a data-driven approach to automatically add new pronunciations to a given phonetic dictionary in a way that they model the given occurrences of words in the database. This algorithm was extended to produce alternative pronunciations for word tuples and frequently "cognized words. Some of the frequently misrecognized words were modeled more accurately by using word tuples and that pronunciations for such tuples were found using Dictionary Learning. The experiments were performed on the GSST, using the speech recognition engine of JANUS 2, the spontaneous speech-to-speech translation system of the Interactive Systems Laboratories at Carnegie Mellon and Karlsruhe University. Experimental results showed that Dictionary Learning algorithm for enhancing and adapting phonetic transcriptions to the existing dictionaries improved the overall recognition performance of the speech recognizer significantly.

Dibazer *et al.* (2004) presented a biologically based speech recognition system. The goal was to develop a noise-robust and feasible speech recognition system, based on fundamental functional principles of the brain. This task was completed in three steps: firstly, speech signal was decomposed into different frequency bands. Secondly, short term energy of signals was encoded into the train of spikes and finally, the classification of temporal patterns was done using dynamic synapse neural networks. The release function of dynamic synapse neural networks was replaced by FD model. The results showed that the performance degradation of dynamic synapse neural networks in the

presence of Gaussian white noise was less than Mel frequency cepstral coefficients. However, the overall recognition rate was less than the mentioned systems.

Deng *et al.* (2007) proposed a novel voiced-unvoiced-silence detection based on unsupervised learning. The class-dependent statistics (feature means, covariance matrices, and occurrence frequencies of voiced, unvoiced, and silence classes) needed for the classification were estimated directly from the signal to be classified via Gaussian mixture models and the expectation maximization algorithm. The classification was tested using NTIMIT database and the accuracy was very near to that of a fully trained classification. Accuracy was greater than 91.15%, and voice activity detection accuracy was greater than 97.45%.

Rapp (2008) described basic aspects and concepts of language modeling for applications in real world automatic speech recognition systems, which are intended to be used by lawyer's chambers, judiciary and law enforcements. He presented only some basic concepts and ideas. Most of the experiments were conducted using HTK toolkit which is reliable, efficient and proven framework for designing and building automatic speech recognition systems.

Sheikan *et al.* (2010) extracted a set of basic features from speech signal by using Haar wavelet transform and then classified them using SVM algorithm. The features extracted were - zero crossing rate, standard deviation and average magnitude. For feature extraction, speech signal was divided into five sub-layers using Haar wavelet and then the mentioned features were extracted and classified using SVM. Two methods were used in classification: one versus of the rest and pair-wise (couple). It was found that the correct classification rate of test data was about 89% when using pair-wise method. This rate decreased to 67% for one versus of the rest method.

Mitra *et al.* (2011) proposed an algorithm for extracting the articulatory information from the speech signal. Firstly, articulatory information in the form of vocal tract constriction variables from the Aurora-2 speech corpus using a neural network based speech-inversion model was estimated. Then, word recognition tasks were performed both for noisy and clean speech using articulatory information in conjunction with traditional

acoustic features. Results showed that word recognition rates were significantly improved. Results suggested that TVs, if estimated properly, can contribute in improving the noise robustness of automatic speech recognition systems. The improvement in the recognition accuracies by incorporating articulatory information in the form of TVs indicated that the acoustic features and TVs provide partially complementary information about speech; neither of them alone provided better accuracy than when both used together. The recognition accuracy improvement obtained as a result of using TVs in addition to the acoustic features was confirmed at a significance level of 0.01%.

Dua *et al*. (2012) presented automatic speech recognition system for isolated words for Punjabi language. HTK toolkit was used to develop the system. From eight speakers data was collected for training of 115 distinct Punjabi words and some samples were collected from six speakers in real time environment for analyzing the performance of system. A GUI was also implemented using java language to make system more interactive. Data was recorded using audacity recording tool. The performance of system was analyzed in two cases: same speakers involved in both training and testing and speakers involved in testing only. It was observed that the system showed its average performance in range of 94% to 96%.

Izzad *et al.* (2013) presented an algorithm to detect speech and non-speech segments in Malay language for spontaneous speech. Three audio features were used that were energy, zero crossing rate and fundamental frequency for the speech/non-speech detection as each feature has unique properties. Experiments were conducted on one-hour Malay language spontaneous speech consisting of more than 20,000 speech/non-speech segments. Results showed that the proposed method achieved 97.8% accuracy rate.

Tamura *et al.* (2014) developed an audio-visual speech recognition interface for mobile devices such as tablet computers and smart cell phones in various environments. In order to test the recognizer and investigate issues related to audio-visual processing on mobile computers, speech data and lip images of 16 subjects in eight conditions were collected, where there were various audio noises and visual difficulties. Then, audio-only speech recognition and visual-only lip reading were conducted. Through these experiments,

some issues were found about data collection. In the speech recognition experiment, the accuracy could not reach 40% in the five environments.

Mankala *et al.* (2014) proposed automatic speech recognizer for Telugu language using HTK toolkit. This was developed for recognizing isolated words using acoustic word model. Data was collected from nine Telugu speakers for training purpose and system was trained using 113 isolated Telugu words. The overall accuracy of the system that was observed was in the range of 95.46% and 96.64%.

2.2 Literature Survey on Prosody Labeling

Xijun *et al.* (2003) presented an automatic prosody labeling system, in which the decision tree was used. In this system, not only the acoustic parameters but also the text information such as the part-of-speech of a word was used. 4 hierarchical layers were used to describe the prosodic structure. A prosody model was built up using the automatically labeled corpus for Mandarin Text To Speech system. Listening test showed that the system worked well. It was an attempt to speed up the corpus building process.

Ananthakrishnan *et al.* (2005) described an automatic prosody recognition system that detected stress and prosodic boundaries at the word and syllable level and used coupled HMMs to model multiple, asynchronous acoustic feature streams and a syntactic-prosodic model that captured the relationship between the syntax of the utterance and its prosodic structure. Results showed that the recognizer achieved about 75% agreement on stress labeling and 88% on boundary labeling at the syllable level.

Chiang *et al.* (2007) proposed a latent prosody model of Mandarin speech for the interaction of tone and prosody state. The main purpose was automatic prosody state labeling and to improve the tone recognition accuracy. Experimental results on the Tree-Bank corpus showed that latent prosody model gave meaningful prosody state labeling results and also improved the average tone recognition rate from 80.86% to 82.55% as compared to the multi-layer perceptron based baseline.

Chongjia *et al.* (2008) presented an automatic prosody boundary label system, based on large speech corpus with prosodic structure label. CART framework was used to classify

prosody boundary. In this system, both the acoustic parameters and the text information were used. It was found that this model achieved an accuracy of 90.86% for prosody boundary detection. It was an attempt to speed up the corpus building process.

Ananthakrishnan *et al.* (2008) proposed a novel unsupervised adaptation approach for improving the quality of prosodic language model and then evaluated it on pitch detection task. Different parts of the adaptation set were weighted according to the confidence level assigned by the seed models during automatic prosody labeling. This weighted data was used to adapt the seed prosodic language model. This algorithm resulted in a relative improvement of 13.8% over the seed prosody language model on the binary pitch accent detection task.

Qian *et al*. (2010) proposed an approach for automatic prosody prediction and detection. It made use of Conditional Random Field with rich, phonetic, syntactic and acoustic, contextual features. Experimental results performed on Boston University Radio Speech Corpus showed that the accuracy of prosody prediction and detection was improved in speaker-dependent as well as speaker independent cases. The prosody model achieved an accuracy of 82.93% and 92.11% for pitch accent and break detection, respectively.

Wang *et al.* (2011) described the design and development of a database of functional and emotional intonation in Chinese. The database was based on conversations from movies and TV plays and duration was about 110 hours. Utterances were segmented, syllable and prosody labeling was done, pitch was extracted and then manually corrected. This database was useful for studying functional and emotional intonation, as well as functional and emotional recognition. It contained a large number of utterances, their transcriptions, and pitch values by different speakers.

Sin-Horng *et al.* (2012) presented a new prosody-assisted automatic speech recognition system for Mandarin speech. The system used a sophisticated prosody modeling method to generate 12 prosodic models. The proposed system had many advantages. Unlabeled speech database was used to train the 12 prosodic models. Experimental results showed that the system significantly performed better than the baseline scheme using an HMM recognizer. Performances of 20.7%, 14.4%, and 9.6% in word, character, and base-

syllable error rates were obtained. By an error analysis, it was found that many word segmentation errors and tone recognition errors were corrected.

Ning *et al*. (2012) made preliminary study for the Tibetan speech prosody. 20 passages each of male and female were collected of Tibetan Lhasa dialect news as subject. They were labeled by Praat software, extracted and then saved into .xls files. Labeling was done for various levels of prosody units which included sentence, prosody-phrase, prosody-word and syllable. After data extraction, interrelationship, different levels of prosody units' duration, silence segmentations, etc. between male and female news reading aspects were studied.

Brognaux *et al.* (2012) proposed a method to detect and correct the amount of labeling errors. It could be applied as a post-process to any syntax-driven prosody labeling. Acoustic information was used to check the syntax-based labels and then, determine those which need to be changed. The proposed method did not require manually prosody-labelled corpus. The evaluation on a corpus in French showed that more than 75% of the errors detected by the method were effective errors which must be corrected.

Sreejith *et al.* (2013) described an approach to automatic labeling of prosodic events. A baseline phonetic engine was created. Data was collected in three different modes - read speech, extempore speech and conversational speech from various regions of Kerala. 5 hours data was collected in each mode from 10 different speakers. An algorithm was proposed to incorporate the prosodic information into baseline system. Automatic pitch marking and break index marking were done in order to represent the prosody.

Deekshitha *et al.* (2015) created a baseline phonetic engine using Malayalam speech database. A GUI was developed for the phonetic engine. Prosodic knowledge was incorporated for the improvement of the baseline phonetic engine. Automatic break index marking and pitch marking were done for prosody representation, and these labels were used for modifying phonetic engine. The developed phonetic engine could be used for finding composition of phoneme in spoken utterance. It was created with sufficient accuracy even in case of long speech utterances by incorporating prosodic breaks.

2.3 Literature Survey on Pitch Marking

Ananthapadmanabha *et al.* (1975) proposed a general theory of epoch extraction of overlapping non-identical waveforms. The theory was applied to outputs of models of voiced speech production mechanism and to actual speech data. Some typical glottal waveshapes were considered to explain their effect on the speech output. For finding the glottal waveform, inverse filtering technique was used. Only an approximate shape of the glottal wave could be estimated by this approach, but the important characteristics of the glottal wave and the variations in pitch period were still obtained. Accurate pitch information could be obtained by this method even for high-pitched sounds.

Ohmura (1994) presented a method for fine pitch contour extraction. In this method, pitch period was determined pitch-synchronously using a simple wave filtering technique to extract the fundamental frequency component of voice. It included two stages. The first stage of processing was wave filtering using an adaptively controlled filter to extract the pitch channel signal. The second stage was the determination process of pitch contour. This algorithm showed very good agreement between the resultant pitch contours obtained with this method and the manually traced standard pitch contours. The results indicated that this method could successfully carry out faithful pitch tracking. The advantages of this method were reliable voice fundamental wave synchronous pitch detection and amplitude information extraction. There were still technical problems with this method to speech in the absence of pitch channel signals.

Harbeck *et al.* (1995) developed a new pitch synchronous algorithm. It interpreted the search for the pitch periods as an optimization problem. In this algorithm, the path through a search space of pitch period hypothesis was determined. The search for the best path was efficiently implemented by dynamic programming. The dynamic programming cost function was determined by artificial neural networks which were automatically trained. It combines the outputs of heuristic functions and measures the similarity between adjacent pitch periods. This algorithm was divided into four steps – preprocessing, zero crossings determination, search space of pitch period hypotheses generation and finally the best path computation. For accuracy computation, the pitch

level which was obtained was compared to the reference pitch level which was available for the speech signal. An error rate of 4.75% was found on German speech database with a deviation factor of 45. The algorithm defined was not computationally efficient due to the application of artificial neural networks and also dynamic programming for computing the minimum cost path.

Goncharoff *et al.* (1998) presented two algorithms – one for estimating the pitch period and another for estimating pitch phase. In both the algorithms, dynamic programming was widely used. In these algorithms, the whole signal was pitch marked because here, unvoiced and voiced intervals were treated equally. A short time energy contour was obtained. This computed energy contour had large amplitude peaks in the voiced regions which were equally spaced and low amplitude peaks in the unvoiced regions at irregular intervals. Dynamic programming was used to compute the most likely pitch pulse location for each frame. This method performed well but no experimental results were given. This method is expected to work better, if the pitch tracking is accurate and the pitch periods do not deviate from cycle-to-cycle.

Mann *et al.* (1998) proposed a novel nonlinear algorithm for epoch marking in voiced speech. The algorithm operated entirely in state space and used Poincare sections for marking only the pitch synchronous points in speech signal. The proposed algorithm worked accurately on simple vowel sounds and rising pitch vowels. It accurately marked all the epochs. Moderate success was achieved for real voiced speech signals.

Laprie *et al*. (1998) proposed an automatic pitch marking algorithm which could be used for modifying speech signals with TD-PSOLA. The algorithm tried to optimize the propagation of pitch marks using the known pitch values obtained from the pitch extraction algorithm. Firstly, all the extrema were extracted in equally spaced voiced regions. Then, using dynamic programming, an optimal set of extremas was found. This algorithm gave satisfied results with all speech signals and also performed well on both male and female speakers. It could easily be combined with any algorithm for pitch extraction.

Colotte *et al.* (2000) presented an approach to slow down the speech signals selectively and to enhance speech intelligibility. Selective slowing down used the TD-PSOLA method. An automatic pitch marking algorithm was designed to apply this method automatically. The enhancement consisted of amplifying stop bursts and unvoiced fricatives. Results showed that the oral comprehension was improved through this approach. The combination of slowing down selectively and acoustic enhancement of stops and fricatives improved the speech intelligibility.

Sakamoto *et al.* (2000) presented an automatic pitch-marking method using the wavelet transform. This method detected discontinuity in the speech waveform which occurred at the glottal closure instant. Experimental results showed that it achieved 96% detection accuracy.

Veldhuis (2000) proposed a pitch marking technique which made use of dynamic programming. Three consistency requirements were suggested for determining candidate pitch markers and also as dynamic programming cost function. These consistency requirements were characteristic property requirement, waveform consistency requirement and pitch consistency requirement. All these three consistency requirements function differently. The waveform consistency requirement checked for the matches between the signal portions around the adjacent pitch markers. Third one, i.e. the pitch consistency requirement selected those markers which were in the vicinity of the estimated pitch period. It was found experimentally that the characteristic property requirement did not strongly influence the accuracy of pitch markers. Moreover, it was found that the marking obtained, without the use of pitch and waveform consistency requirements did not have much accuracy. This algorithm was also sensitive to large errors in estimated pitch periods.

Colotte *et al.* (2002) proposed an algorithm for higher precision pitch marking with TD-PSOLA. The main objective was to preserve the consistency of marks between the neighbouring frames with respect to the temporal structure of pitch periods. First of all, pitch marking was improved by eliminating the mismatch errors which appeared during rapid formant transitions. Then, this improved pitch marking was combined with a fast

resampling technique during synthesis step to increase the quality of signal and achieve a higher precision synthesis with TD-PSOLA.

Ashouri *et al.* (2004) developed an expert system based on a new pitch marking algorithm and three already existing pitch marking tools. This expert system used simple logical combinations of these tools outputs. The behaviour of human expert system was taken into account to develop the post-processing which was necessary to complete each tool and to improve the results of their combinations. However, accurate and complete pitch marking was best achieved with all the four outputs at the expense of some higher processing time. The developed system was trained and tested using a data-base of Farsi vowels and semi-vowels.

Chalamandaris *et al.* (2009) proposed an approach to calculate the pitch marks of a speech signal in a specific framework of TTS system. Using dynamic programming, it identified the analysis pitch marks accurately and efficiently for a TD-PSOLA TTS system. It helped the TTS system to provide a high quality synthetic speech with low signal distortion. Experimental results showed that this algorithm worked better with the TD-PSOLA-based TTS synthesizer as compared to many other algorithms.

Alias *et al*. (2010) introduced a methodology to achieve reliable pitch marking on the affective speech. A three-stage restricted dynamic programming was applied to adjust the pitch marks at signal peaks or valleys. This approach could be applied as a post-processing of any pitch determination or pitch marking algorithm, or their merging. The experiments showed that the proposed approach significantly improved the results of the input state-of-the-art markers on affective speech. This algorithm gave significant improvement for any RDP configuration, input marker, and speaking style.

Chang *et al*. (2012) proposed a robust, real time and accurate pitch determination algorithm for telephone speech. This algorithm used original as well as the absolute speech signal to calculate a combined Auto Correlation Function. A non-linear process Pitch Candidate Refinement Function that incorporates Long-Time Average Pitch was used to refine the combined ACF. Then, based on combined ACF, the pitch candidates were selected and dynamic programming was used to track down to the true pitch.

Experimental results on Keele pitch database showed that this algorithm can reduce the gross error rate for telephone speech by about 3%. All the operations were in time domain, so the algorithm was easy to implement and time efficient.

Buza *et al.* (2013) presented an algorithm to detect the pitch intervals of voice signal. It worked in four steps where each step needed a separate algorithm. Firstly, pivot point was determined, secondly, an estimate of the period around pivot point was determined, then, glottal peaks and hiatus points were determined and finally, end points of pitch intervals were determined. The proposed algorithm was very fast, efficient and accurate as it worked exclusively in the time domain. It performed well on voiced speech segments.

Ykhlef *et al.* (2013) proposed a new pitch marking method to locate the Global Peak Instants of voiced speech regions. It was based on the speech waveform to extract the pitch marks. It used the mean based signal to locate the time intervals from where the global peak instants were extracted. This method did not need any pitch detection algorithm or LP residual information. Results showed that the obtained pitch marks were correct and allowed pitch and duration modifications of speech with good quality. The main drawback of this method was the false alarm and missed marks that occured in some regions of speech. So, for performance improvement, a post processing stage was added to reduce the erroneous pitch marks.

CHAPTER 3

Problem Statement, Data Collection and Preparation

3.1 Problem Statement

Automatic prosody labeling through pitch marking is the main objective of this thesis. Prosody labeling is done to describe the prosodic events: pitch and duration. It is the most important content in linguistic research and voice projects. Incorporating prosodic knowledge into automatic speech recognition improves the performance of the phonetic engine. Application of prosodic features in speech processing is already established by various people but unfortunately, incorporating prosody into speech systems has to face various challenges. These are:

i. Extraction and representation of prosody knowledge.
ii. Labeling of prosodic events manually is very time consuming, laborious and requires several man hours.

So, the need of automatic prosody labeling arises.

The work presented in this thesis focuses on pitch marking, a prosodic event. It is an important task for prosody modifications in speech synthesis. Pitch marks can be defined as the location of a signal period in a voiced segment of speech, labeling the fundamental periodicity of speech. The exact moments of the excitation of vocal tract are determined.

3.2 Data Collection and Preparation

For the work presented in this thesis, data is collected in two different modes: read speech mode and lecture mode.

i. Read Speech Mode :

In this mode, data is collected from seven different Punjabi speakers (2 male and 5 female members) for a duration of about 22 minutes. Four different recordings

are taken for each speaker. So, total number of recordings of read speech comes out to be 28. Duration of each recording lies in the interval - 24 seconds to 61 seconds.

ii. Lecture Mode :

In lecture mode, data is collected from three different Punjabi speakers for a duration of about 10 minutes. Here also, 4 different recordings are taken of each speaker. So, total number of recordings in case of lecture mode comes out to be 12. Duration of each recording lies in the interval - 46 seconds to 57 seconds.

The total duration of data collected for both modes is shown in Table 1.

Table 1: Total duration of each mode of speech

Mode	No. of Recordings	Duration (in Minutes)
Read Speech Mode	28	22.09
Lecture Mode	12	10.04

So, the total data collected consists of 40 different recordings and the total duration is 32 minutes and 13 seconds. All the recordings were cut from files of long duration and then saved in wav file format.

Pitch marking algorithm, discussed in the next chapter is applied to all the 40 files. In addition, 4 different users manually did the pitch marking of these 40 files. Automatic labeling of prosody through pitch marking is done and then compared to the manual pitch marking.

CHAPTER 4

Algorithm for Pitch Marking, Results and Discussion

4.1 Algorithm for Pitch Marking

4.1.1 Introduction

The main objective of this thesis is to present a method for pitch accent marking which labels variations in pitch. The source code has been taken from the consortium member working in the project entitled "Automation of Pitch Accent Markings in Punjabi Speech". It has been implemented in MATLAB with some modifications made to it. The complete algorithm is discussed in next section. A GUI has also been developed which records a voice and then, does its pitch marking. The GUI has been shown in Appendix.

4.1.2 Algorithm

The steps followed for the automation of pitch accent markings are as follows:

Step 1: An audio voice is recorded and then saved in wav file format.

Step 2: The sampling rate is checked for this audio file. For pitch marking, it is mandatory that the sampling rate of wave file should not be more than 8000 Hz. So, ensure that the sampling rate is not more than 8000 Hz using resample().

Step 3: A .pt file gets created after running the algorithm. This file contains the pitch markings corresponding to their time intervals.

A sample .pt file is shown in Figure 7.

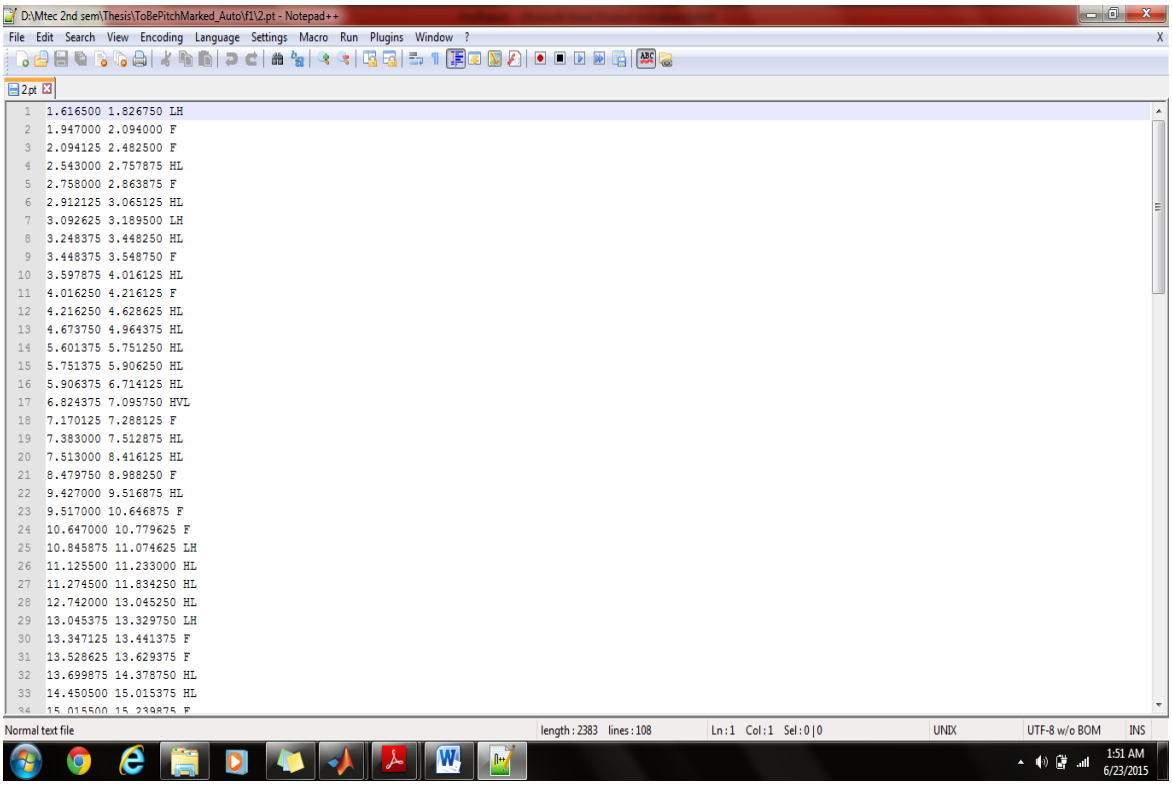

Figure 7: Sample .pt file

Step 4: Noise is added to the signal to return the final noise added signal. For this, generated noise power and input signal power are calculated.

Step 5: Then, instances of glottal closure are found.

Step 6: Remove spurious by epoch drift and pitch period information. Only some instants of glottal closure are stored. Spurious part is also removed by applying some other conditions.

Step 7: Voiced-unvoiced boundary marking is done as pitch works only with the voiced speech. No pitch periods in unvoiced speech.

Step 8: Pitch marking is then done based on the pitch frequency.

Step 9: Finally, the output is written to the wave file.

A value of 1 is given to the F region, value = 2 is given to the LH region, value = 3 given to the HL region, value = 3.7 given to the VHL region, value = 3.5 given to the HVL region, value = 2.7 given to the VLH region and value = 2.5 given to the LVH region.

4.2 Results and Discussion

After applying the algorithm described above, total 4 figures are obtained which are discussed in section 4.2.1. Section 4.2.2 shows the system generated pitch marking on WaveSurfer. Section 4.2.3 shows the same file pitch marked manually and the comparison between system generated and manual pitch marking. All these results are noted on an audio file of 4 seconds duration. Finally, section 4.2.4 shows the results of the algorithm and manual pitch marking on 40 different audio files.

4.2.1 Outputs of Algorithm

Figure 8 shows the audio signal which is considered for pitch marking.

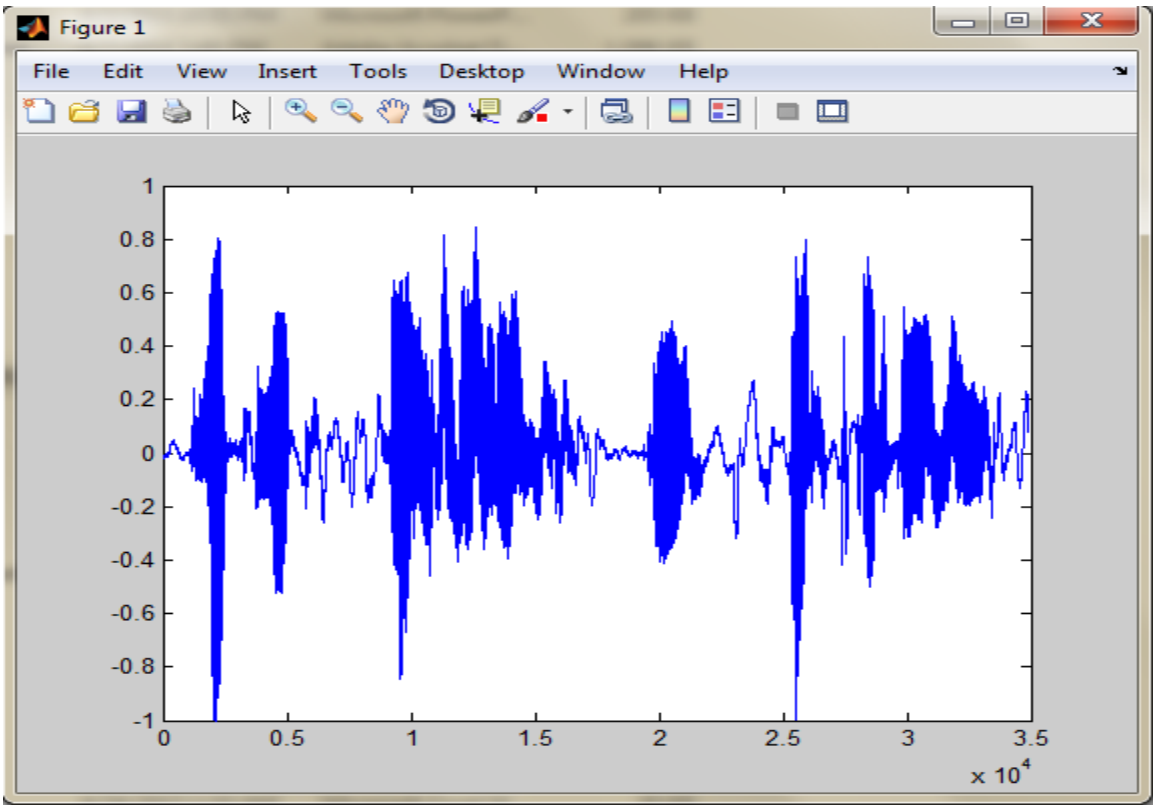

Figure 8: Audio signal

Figure 9 depicts the voiced regions marked as red.

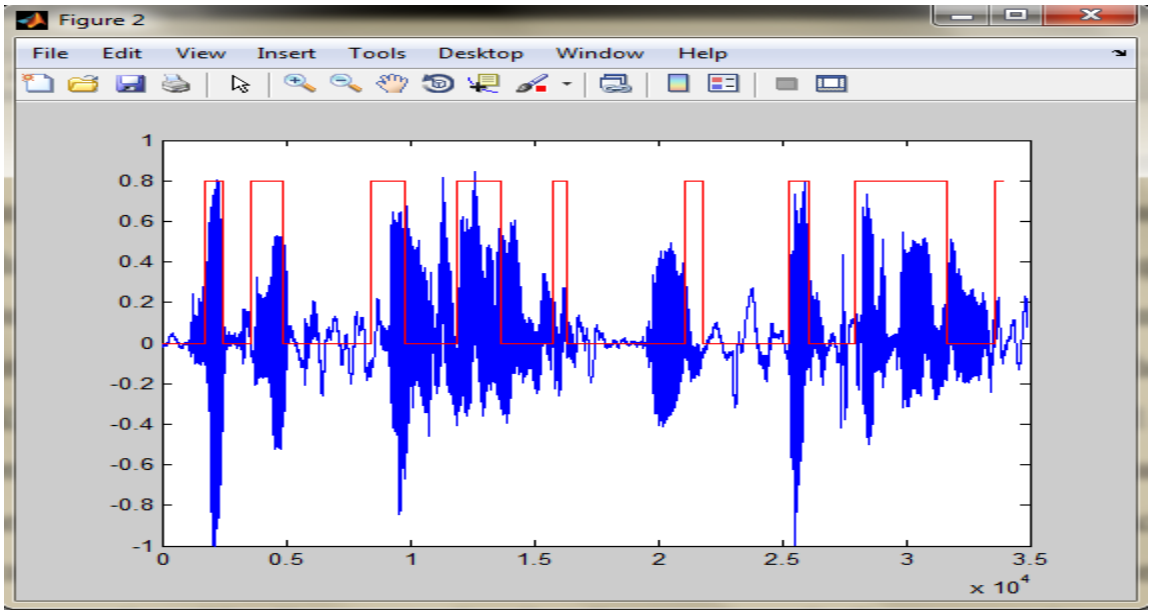

Figure 9: Voiced regions

Figure 10 depicts the starting and ending of voiced region with 'Red' with 'Black' respectively.

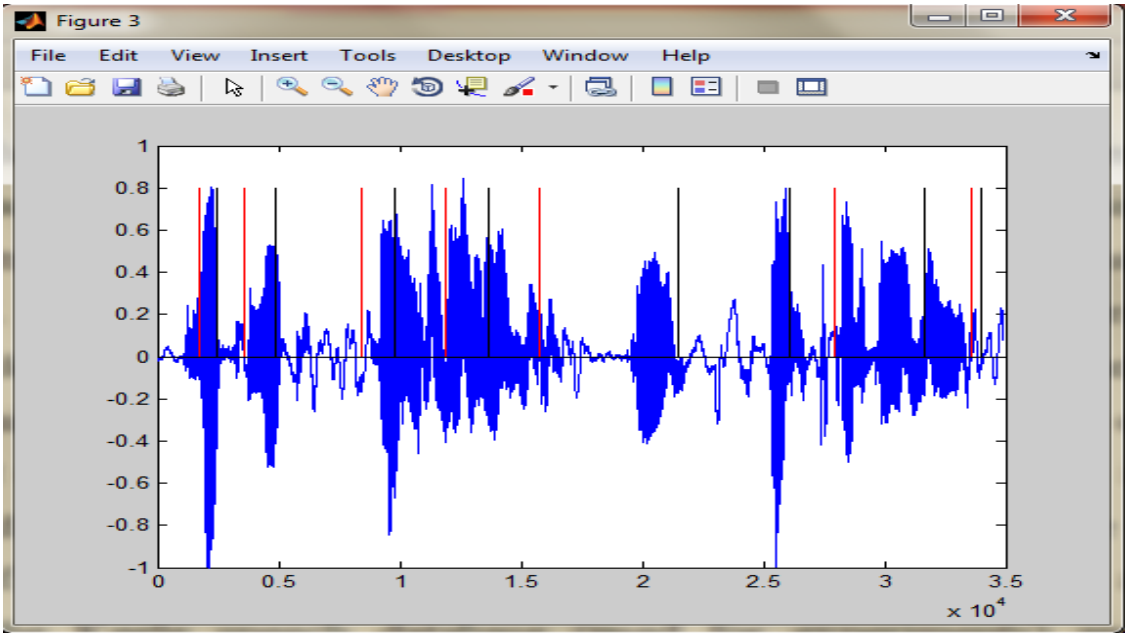

Figure 10: Starting of voiced region with 'Red' and ending with 'Black'

Figure 11 shows the final pitch marking.

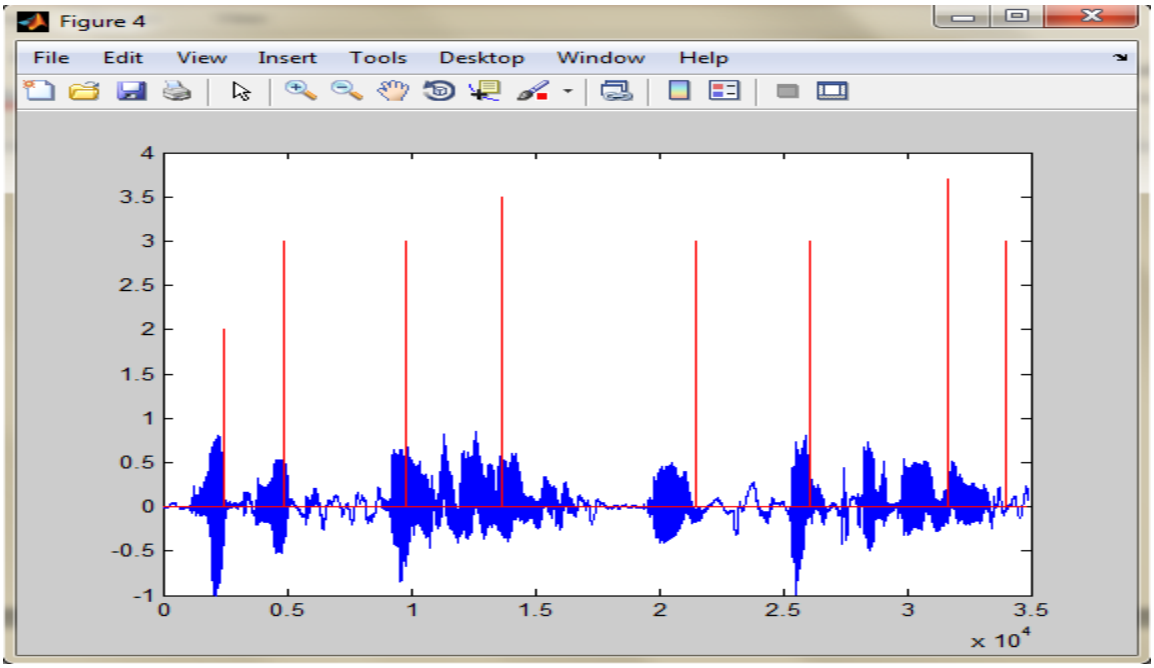

Figure 11: Pitch marking

4.2.2 Output on WaveSurfer

Figure 12 shows the system generated pitch marking on WaveSurfer.

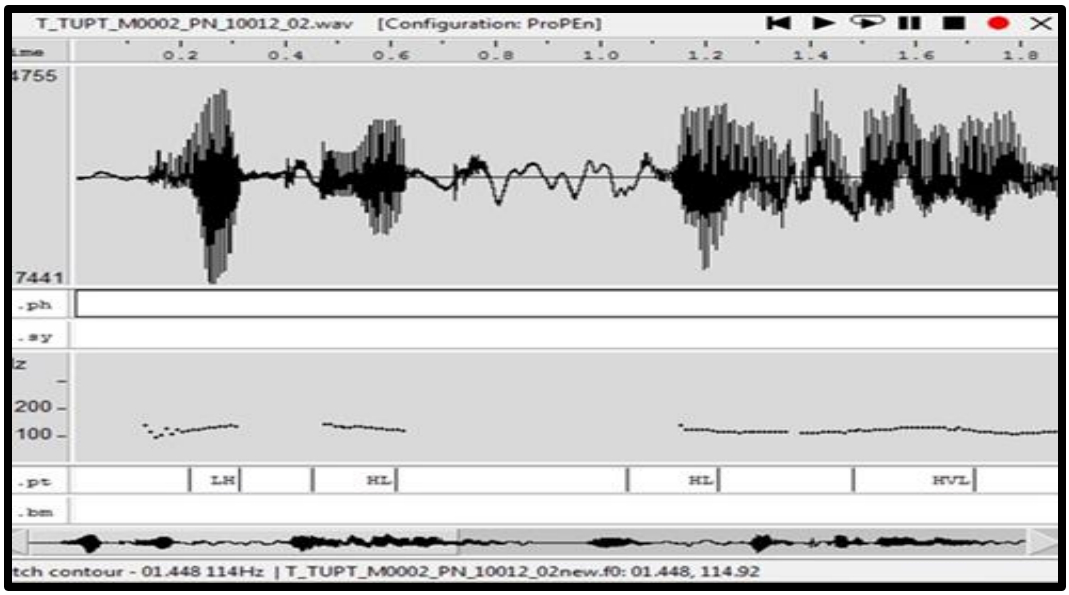

Figure 12: System generated pitch marking

It can be clearly seen from Figure 11 and Figure 12 that a value of 2 represents Flat region, value of 3 represents HL, 3.5 represents HVL and so on.

4.2.3 Manual Pitch Marking

Manual pitch marking of the same file is being done and shown in Figure 13.

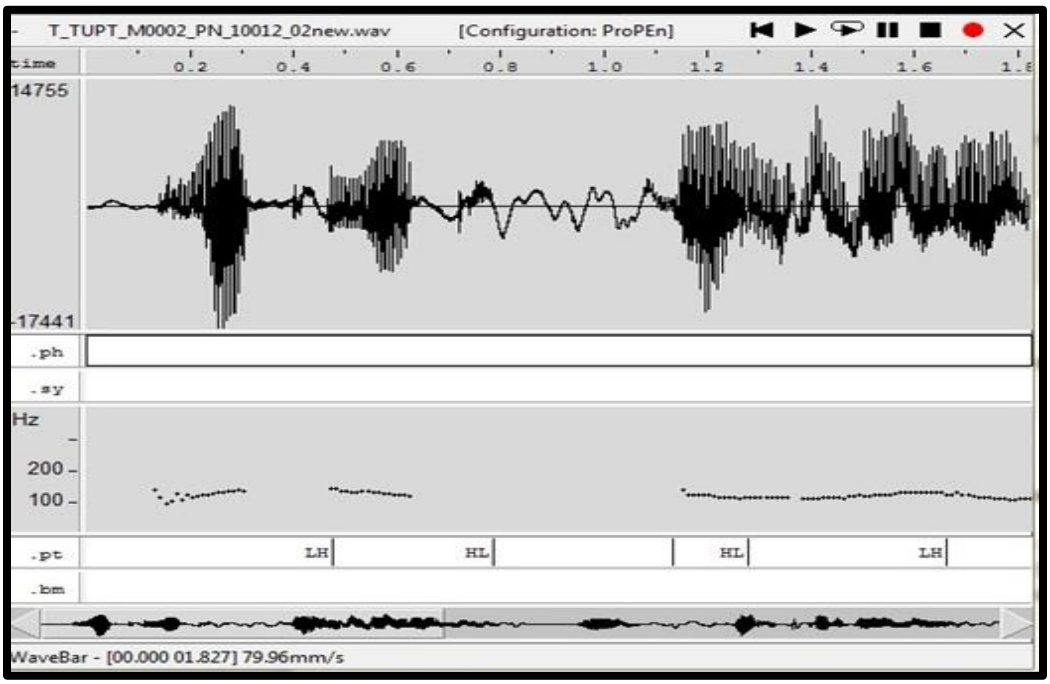

Figure 13: Manual pitch marking

Comparing both system generated and manual pitch marking (Figure 12 and Figure 13), it has been noted that LH segment appeared 4 times if we do the marking manually. In the case of automation, this segment appears only once, i.e., it has an error of 3 occurrences. Similarly the occurrences of other markings have been noted and reported in Table 1. All these results apply only to that particular file.

Table 2: Results of manual vs. system generated pitch marking

	LH	HL	F	VLH	VHL	LVH	HVL
Manual	4	3	1	0	1	0	0
Automated	1	5	0	0	1	0	1
Error	3	2	1	0	0	0	1

4.2.4 Results of 40 Files

This section shows the results of the algorithm on the 40 audio files discussed in Chapter 3. Also, 4 different Human Evaluators: Human Evaluator1 (HE1), Human Evaluator2 (HE2), Human Evaluator3 (HE3) and Human Evaluator4 (HE4) manually pitch marked each file. The number of occurrences of F, HL, LH, VHL, HVL, VLH, LVH in case of automated and manual marking is shown in form of tables along with the duration of each file. Now, the error between the automated and each HE is computed and finally, the average error is calculated for each segment of the file. Tables 3-6 present this data for Speaker 1; Tables 7-10 present this data for Speaker 2; Tables 11-14 present this data for Speaker 3; Tables 15-18 present this data for Speaker 4; Tables 19-22 present this data for Speaker 5; Tables 23-26 present this data for Speaker 6; Tables 27-30 present this data for Speaker 7; Tables 31-34 present this data for Speaker 8; Tables 35-38 present this data for Speaker 9 and Tables 39-42 present this data for Speaker 10.

Speaker 1; File duration: 45 seconds

Table 3: Results of manual vs. system generated pitch marking for 1st file

	Automated	**HE1**	**Error HE1**	**HE2**	**Error HE2**	**HE3**	**Error HE3**	**HE4**	**Error HE4**	**Average Error**
F	42	36	6	44	2	8	34	27	15	14.25
HL	41	56	15	74	33	79	38	39	2	22
LH	15	20	5	58	43	44	29	33	18	23.75
VHL	3	7	4	1	2	5	2	9	6	3.5
HVL	2	2	0	10	8	0	2	5	3	3.25
VLH	2	0	2	2	0	2	0	0	2	1
LVH	2	4	2	9	7	2	0	15	13	5.5

Speaker 1; File duration: 59 seconds

Table 4: Results of manual vs. system generated pitch marking for 2nd file

	Automated	**HE1**	**Error HE1**	**HE2**	**Error HE2**	**HE3**	**Error HE3**	**HE4**	**Error HE4**	**Average Error**
F	54	43	11	39	15	16	38	22	32	24
HL	58	69	11	82	24	91	33	53	5	18.25
LH	22	25	3	77	55	43	21	34	12	22.75
VHL	1	12	11	5	4	2	1	21	20	9
HVL	5	1	4	18	13	3	2	6	1	5
VLH	3	0	3	5	2	0	3	0	3	2.75
LVH	1	3	2	3	2	4	3	21	20	6.75

Speaker 1; File duration: 53 seconds

Table 5: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	57	40	17	27	30	9	48	10	47	35.5
HL	50	74	24	102	52	121	71	36	14	40.25
LH	25	28	3	75	50	65	40	27	2	23.75
VHL	0	8	8	7	7	4	4	18	18	9.25
HVL	4	0	4	13	9	2	2	4	0	3.75
VLH	2	0	2	0	2	3	1	0	2	1.75
LVH	1	11	10	11	10	3	2	24	23	11.25

Speaker 1; File duration: 45 seconds

Table 6: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	45	35	10	18	27	7	38	34	11	21.5
HL	31	38	7	81	50	76	45	37	6	27
LH	28	26	2	53	25	26	2	29	1	7.5
VHL	1	3	2	2	1	1	0	13	12	3.75
HVL	1	2	1	7	6	14	13	3	2	5.5
VLH	2	0	2	1	1	3	1	0	2	1.5
LVH	5	8	3	5	0	1	4	25	20	6.75

It can clearly be observed from Tables 3-6 (Speaker 1) that the average error for F, HL and LH segment is very high as compared to the other four. VLH segment has the least average error.

Speaker 2; File duration: 51 seconds

Table 7: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	36	38	2	18	18	15	21	45	9	12.5
HL	45	62	17	93	48	66	21	87	42	32
LH	20	26	6	40	20	16	4	71	51	20.25
VHL	4	4	0	5	1	1	3	45	41	11.25
HVL	8	3	5	10	2	17	9	7	1	4.25
VLH	4	0	4	1	3	2	2	3	1	2.5
LVH	2	1	1	1	1	0	2	40	38	10.5

Speaker 2; File duration: 52 seconds

Table 8: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	53	27	26	24	29	12	41	47	6	25.5
HL	34	67	33	92	58	102	68	100	66	56.25
LH	26	28	2	38	12	27	1	59	33	12
VHL	2	1	1	5	3	0	2	14	12	12
HVL	6	0	6	7	1	5	1	7	1	2.25
VLH	6	0	6	0	6	0	6	3	3	5.25
LVH	0	3	3	1	1	0	0	25	25	7.25

Speaker 2; File duration: 49 seconds

Table 9: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	38	28	10	22	16	5	33	37	1	15
HL	40	51	11	72	32	66	26	86	46	28.75
LH	23	39	16	39	16	27	4	71	48	21
VHL	7	2	5	7	0	1	6	21	14	6.25
HVL	7	0	7	4	3	8	1	6	1	3
VLH	2	0	2	0	2	0	2	1	1	1.75
LVH	4	1	3	1	3	0	4	29	25	8.75

Speaker 2; File duration: 46 seconds

Table 10: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	26	26	0	13	13	6	20	20	6	9.75
HL	53	52	1	72	19	63	10	58	5	8.75
LH	28	42	14	40	12	34	6	55	27	14.75
VHL	9	1	8	6	3	6	3	18	9	5.75
HVL	11	2	9	1	10	4	7	8	3	7.25
VLH	3	0	3	2	1	0	3	6	3	2.5
LVH	5	4	1	0	5	7	2	22	17	6.25

It can clearly be observed from Tables 7-10 (Speaker 2) that the average error for F, HL and LH segment is very high as compared to the other four. HL segment has the maximum average error except for the last file and VLH has the least average error except for the 2nd file.

Speaker 3; File duration: 24 seconds

Table 11: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	26	19	7	16	10	1	25	11	15	14.25
HL	28	23	5	45	17	53	25	48	20	16.75
LH	13	21	8	20	7	13	0	37	24	9.75
VHL	1	2	1	1	0	0	1	17	16	4.5
HVL	5	4	1	5	0	2	3	8	3	1.75
VLH	2	0	2	0	2	0	2	2	0	1.5
LVH	0	2	2	0	0	0	0	12	12	3.5

Speaker 3; File duration: 38 seconds

Table 12: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	21	17	4	16	5	1	20	33	12	10.25
HL	32	34	2	50	18	87	55	65	33	27
LH	22	32	10	29	7	24	2	47	25	11
VHL	2	3	1	1	1	0	2	25	23	6.75
HVL	5	2	3	8	3	1	4	14	9	4.75
VLH	3	0	3	0	3	0	3	3	0	2.25
LVH	3	3	0	5	2	0	3	26	23	7

Speaker 3; File duration: 51 seconds

Table 13: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	32	27	5	10	22	5	27	47	15	17.25
HL	50	46	4	79	29	117	67	100	50	37.5
LH	45	33	12	27	18	29	16	50	5	12.75
VHL	4	4	0	1	3	0	4	38	34	10.25
HVL	7	2	5	5	2	1	6	15	8	5.25
VLH	2	0	2	0	2	0	2	4	2	2
LVH	8	4	4	9	1	0	8	53	45	14.5

Speaker 3; File duration: 42 seconds

Table 14: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	23	22	1	7	16	2	21	25	2	10
HL	32	36	4	72	40	81	49	76	44	34.25
LH	35	26	9	23	12	25	10	54	19	12.5
VHL	4	5	1	3	1	0	4	24	20	6.5
HVL	4	1	3	4	0	1	3	15	11	4.25
VLH	6	0	6	0	6	0	6	3	3	5.25
LVH	3	5	2	0	3	0	3	31	28	9

It can clearly be observed from Tables 11-14 (Speaker 3) that the average error for F, HL and LH segment is comparatively higher than the other four. HL segment has the maximum average error and VLH has the least average error.

Speaker 4; File duration: 55 seconds

Table 15: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	38	19	19	13	25	4	34	48	10	22
HL	39	30	9	56	17	65	26	108	69	30.25
LH	10	27	17	32	22	23	13	91	81	33.25
VHL	7	9	2	3	4	0	7	13	6	4.75
HVL	0	3	3	4	4	5	5	8	8	5
VLH	0	0	0	0	0	0	0	1	1	0.25
LVH	2	3	1	1	1	0	2	9	7	2.75

Speaker 4; File duration: 61 seconds

Table 16: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	48	30	18	25	23	1	47	55	7	23.75
HL	48	42	6	71	23	101	53	141	93	43.75
LH	15	35	20	24	9	19	4	106	91	31
VHL	9	6	3	1	8	0	9	8	1	5.25
HVL	0	2	2	5	5	2	2	22	22	7.75
VLH	1	0	1	1	0	0	1	0	1	0.75
LVH	3	6	3	1	2	0	3	7	4	3

Speaker 4; File duration: 50 seconds

Table 17: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	46	30	16	23	23	6	40	41	5	21
HL	36	40	4	79	43	77	41	119	83	42.75
LH	16	29	13	15	1	18	2	78	62	19.5
VHL	2	4	2	3	1	3	1	6	4	2
HVL	1	3	2	4	3	3	2	14	13	5
VLH	0	0	0	1	1	0	0	2	2	0.75
LVH	1	3	2	1	0	1	0	17	16	4.5

Speaker 4; File duration: 55 seconds

Table 18: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	36	34	2	19	17	1	35	55	19	18.25
HL	42	47	5	80	38	84	42	110	68	38.25
LH	15	25	10	23	8	17	2	93	78	24.5
VHL	6	4	2	3	3	3	3	5	1	2.25
HVL	0	2	2	5	5	3	3	12	12	5.5
VLH	1	0	1	0	1	1	0	0	1	0.75
LVH	5	3	2	0	5	0	5	2	3	3.75

It can clearly be observed from Tables 15-18 (Speaker 4) that the average error for F, HL and LH segment is very high as compared to the other four. HL segment has the maximum average error except for the 1st file and VLH has the least average error.

Speaker 5; File duration: 55 seconds

Table 19: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	65	39	26	24	41	6	59	47	18	36
HL	46	64	18	115	69	84	38	136	90	53.75
LH	49	48	1	51	2	34	15	113	64	20.5
VHL	0	4	4	3	3	1	1	0	0	2
HVL	6	3	3	2	4	5	1	10	4	3
VLH	8	3	5	0	8	2	6	0	8	6.75
LVH	0	6	6	3	3	0	0	1	1	2.5

Speaker 5; File duration: 44 seconds

Table 20: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	60	45	15	23	37	4	56	30	30	34.5
HL	28	46	18	92	64	68	40	115	87	52.25
LH	32	43	11	41	9	21	11	102	70	25.25
VHL	0	1	1	4	4	2	2	1	1	2
HVL	1	1	0	2	1	3	2	8	7	2.5
VLH	3	2	1	1	2	0	3	0	3	2.25
LVH	0	2	2	0	0	0	0	2	2	1

Speaker 5; File duration: 50 seconds

Table 21: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	44	42	2	23	21	3	41	28	16	20
HL	66	41	25	109	43	87	21	122	56	36.25
LH	65	58	7	49	16	39	26	91	26	18.75
VHL	0	7	7	5	5	2	2	3	3	4.25
HVL	12	5	7	6	6	4	8	13	1	5.5
VLH	8	0	8	1	7	0	8	1	7	7.5
LVH	0	2	2	0	0	0	0	5	5	1.75

Speaker 5; File duration: 46 seconds

Table 22: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	48	30	18	14	34	3	45	47	1	24.5
HL	41	45	4	113	72	83	42	124	83	50.25
LH	40	53	13	36	4	37	3	105	65	21.25
VHL	0	6	6	8	8	2	2	1	1	4.25
HVL	3	3	0	6	3	2	1	12	9	3.25
VLH	9	3	6	0	9	0	9	1	8	8
LVH	0	8	8	4	4	0	0	3	3	3.75

It can clearly be observed from Tables 19-22 (Speaker 5) that the average error for F, HL and LH segment is very high as compared to the other four. HL segment has the maximum average error.

Speaker 6; File duration: 48 seconds

Table 23: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	37	31	6	19	18	0	37	30	7	17
HL	43	55	12	89	46	64	21	66	23	25.5
LH	37	47	10	46	9	38	1	70	33	13.25
VHL	1	6	5	5	4	0	1	5	4	3.5
HVL	13	2	11	10	3	3	10	10	3	6.75
VLH	5	1	4	8	3	1	4	1	4	3.75
LVH	0	5	5	2	2	0	0	8	8	3.75

Speaker 6; File duration: 54 seconds

Table 24: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	37	26	11	13	24	1	36	21	16	21.75
HL	55	56	1	86	31	90	35	66	11	19.5
LH	34	42	8	60	26	37	3	75	41	19.5
VHL	3	4	1	3	0	0	3	5	2	1.5
HVL	21	5	16	14	7	2	19	9	12	13.5
VLH	5	2	3	2	3	1	4	2	3	3.25
LVH	2	10	8	2	0	0	2	9	7	4.25

Speaker 6; File duration: 46 seconds

Table 25: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	29	32	3	7	22	2	27	20	9	15.25
HL	44	38	6	107	63	87	43	59	15	31.75
LH	52	35	17	34	18	25	27	54	2	16
VHL	0	4	4	1	1	0	0	3	3	2
HVL	17	5	12	17	0	2	15	20	3	7.5
VLH	17	1	16	4	13	2	15	4	13	14.25
LVH	2	10	8	1	1	0	2	8	6	4.25

Speaker 6; File duration: 57 seconds

Table 26: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	41	47	6	20	21	4	37	29	12	19
HL	63	64	1	139	76	111	48	103	40	41.25
LH	54	51	3	47	7	57	3	101	47	15
VHL	2	8	6	1	1	0	2	1	1	2.5
HVL	19	6	13	18	1	0	19	18	1	8.5
VLH	18	0	18	3	15	0	18	1	17	17
LVH	2	3	1	3	1	1	1	6	4	1.75

It can clearly be observed from Tables 23-26 (Speaker 6) that the average error for F, HL and LH segment is comparatively higher than the other four. HL segment has the maximum average error except for the 2nd file and VHL has the least average error.

Speaker 7; File duration: 51 seconds

Table 27: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	44	17	27	6	36	0	44	30	14	30.25
HL	41	44	3	87	46	81	40	94	53	35.5
LH	35	27	8	43	8	36	1	73	38	13.75
VHL	0	9	9	13	13	0	0	7	7	7.25
HVL	2	11	9	10	8	4	2	4	2	5.25
VLH	3	4	1	2	1	0	3	3	0	1.25
LVH	0	15	15	5	5	2	2	12	12	8.5

Speaker 7; File duration: 47 seconds

Table 28: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	38	21	17	9	29	0	38	35	3	21.75
HL	50	49	1	122	72	94	44	104	54	42.75
LH	36	43	7	43	7	33	3	64	28	11.25
VHL	0	15	15	9	9	3	3	8	8	8.75
HVL	4	7	3	2	2	3	1	4	0	1.5
VLH	3	0	3	2	1	1	2	4	1	1.75
LVH	0	11	11	1	1	0	0	5	5	4.25

Speaker 7; File duration: 49 seconds

Table 29: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	19	25	6	6	13	2	17	29	10	11.5
HL	45	37	8	78	33	70	25	88	43	27.25
LH	25	15	10	24	1	29	4	63	38	13.25
VHL	0	16	16	8	8	5	5	3	3	8
HVL	6	10	4	8	2	1	5	16	10	5.25
VLH	5	2	3	2	3	0	5	1	4	3.75
LVH	0	7	7	4	4	1	1	10	10	5.5

Speaker 7; File duration: 49 seconds

Table 30: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	38	27	11	10	28	1	37	31	7	20.75
HL	48	48	0	123	75	95	47	85	37	39.75
LH	33	28	5	29	4	36	3	74	41	13.25
VHL	0	17	17	5	5	3	3	13	13	9.5
HVL	5	1	4	3	2	1	4	3	2	3
VLH	3	2	1	0	3	0	3	1	2	2.25
LVH	1	13	12	3	2	0	1	10	9	6

It can clearly be observed from Tables 27-30 (Speaker 7) that the average error for F, HL and LH segment is very high as compared to the other four. HL segment has the maximum average error and VLH has the least average error.

Speaker 8; File duration: 48 seconds

Table 31: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	21	18	3	3	18	1	20	20	1	10.5
HL	43	30	13	52	9	50	7	44	1	7.5
LH	42	13	29	18	24	7	35	27	15	25.75
VHL	0	6	6	0	0	0	0	2	2	2
HVL	10	2	8	10	0	3	7	8	2	4.25
VLH	12	0	12	0	12	0	12	1	11	11.75
LVH	0	4	4	0	0	0	0	1	1	1.25

Speaker 8; File duration: 46 seconds

Table 32: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	12	16	4	5	7	1	11	26	14	9
HL	40	25	15	65	25	55	15	54	14	17.25
LH	38	14	24	14	24	9	29	29	9	21.5
VHL	2	11	9	4	2	1	1	4	2	3.5
HVL	12	4	8	4	8	3	9	7	5	7.5
VLH	6	0	6	1	5	0	6	2	4	5.25
LVH	0	3	3	0	0	1	1	13	13	4.25

Speaker 8; File duration: 55 seconds

Table 33: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	32	31	1	16	16	1	31	25	7	13.75
HL	67	34	33	108	41	101	34	78	11	29.75
LH	55	35	20	45	10	44	11	71	16	14.25
VHL	9	26	17	14	5	6	3	9	0	6.25
HVL	18	4	14	5	13	2	16	7	11	13.5
VLH	12	1	11	1	11	0	12	1	11	11.25
LVH	4	13	9	0	4	1	3	13	9	6.25

Speaker 8; File duration: 54 seconds

Table 34: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	45	34	11	9	36	0	45	32	13	26.25
HL	50	55	5	109	59	104	54	79	29	36.75
LH	55	23	32	38	17	24	31	66	11	22.75
VHL	3	16	13	7	4	4	1	6	3	5.25
HVL	12	5	7	7	5	2	10	12	0	5.5
VLH	11	0	11	0	11	0	11	0	11	11
LVH	4	8	4	2	2	1	3	12	8	4.25

It can clearly be observed from Tables 31-34 (Speaker 8) that the average error for F, HL and LH segment is high as compared to the other four. LH segment has the maximum average error for 1st and 2nd file whereas HL segment has the maximum average error for 3rd and 4th file. LVH and VHL segments have almost equal average error.

Speaker 9; File duration: 53 seconds

Table 35: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	15	30	15	6	9	1	14	48	33	17.75
HL	31	41	10	76	45	67	36	79	48	34.75
LH	26	23	3	27	1	34	8	78	52	16
VHL	5	6	1	1	4	0	5	5	0	2.5
HVL	13	2	11	2	11	1	12	3	10	11
VLH	5	0	5	0	5	0	5	0	5	5
LVH	7	10	3	2	5	0	7	3	4	4.75

Speaker 9; File duration: 52 seconds

Table 36: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	16	26	10	13	3	1	15	55	39	16.75
HL	36	37	1	58	22	69	33	74	38	23.5
LH	23	32	9	23	0	29	6	66	43	14.5
VHL	3	1	2	0	3	0	3	0	3	2.75
HVL	13	1	12	1	12	2	11	2	11	11.5
VLH	3	1	2	0	3	0	3	1	2	2.5
LVH	3	0	3	0	3	0	3	2	1	2.5

Speaker 9; File duration: 52 seconds

Table 37: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	34	26	8	16	18	0	34	49	15	18.75
HL	26	42	16	70	44	78	52	86	60	43
LH	18	35	17	27	9	33	15	76	58	24.75
VHL	3	1	2	0	3	0	3	0	3	2.75
HVL	8	1	7	1	7	1	7	2	6	6.75
VLH	2	0	2	0	2	0	2	3	1	1.75
LVH	5	0	5	0	5	0	5	2	3	4.5

Speaker 9; File duration: 42 seconds

Table 38: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	22	16	6	6	16	1	21	43	21	16
HL	21	30	9	64	43	66	45	77	56	38.25
LH	12	24	12	18	6	22	10	65	53	20.25
VHL	1	0	1	0	1	0	1	0	1	1
HVL	7	5	2	3	4	0	7	3	4	4.25
VLH	2	1	1	0	2	0	2	0	2	1.75
LVH	5	1	4	0	5	0	5	1	4	4.5

It can clearly be observed from Tables 35-38 (Speaker 9) that the average error for F, HL and LH segment is very high as compared to the other four. HL segment has the maximum average error.

Speaker 10; File duration: 27 seconds

Table 39: Results of manual vs. system generated pitch marking for 1st file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	21	22	1	13	8	5	16	14	7	8
HL	23	46	23	63	40	67	44	75	52	39.75
LH	7	15	8	23	16	19	12	52	45	20.25
VHL	0	4	4	2	2	0	0	21	21	6.75
HVL	5	0	5	1	4	1	4	5	0	3.25
VLH	1	0	1	0	1	0	1	1	0	0.75
LVH	0	3	3	0	0	0	0	18	18	5.25

Speaker 10; File duration: 47 seconds

Table 40: Results of manual vs. system generated pitch marking for 2nd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	47	30	17	12	35	0	47	41	6	26.25
HL	51	58	7	99	48	108	57	137	86	49.5
LH	31	34	3	44	13	35	4	86	55	18.75
VHL	2	22	20	6	4	6	4	10	8	9
HVL	10	0	10	9	1	1	9	11	1	5.25
VLH	6	0	6	1	5	1	5	0	6	5.5
LVH	0	8	8	1	1	1	1	23	23	8.25

Speaker 10; File duration: 42 seconds

Table 41: Results of manual vs. system generated pitch marking for 3rd file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	42	32	10	6	36	1	41	29	13	25
HL	36	56	20	105	69	97	61	118	82	58
LH	24	24	0	41	17	30	6	75	51	18.5
VHL	1	13	12	2	1	6	5	4	3	5.25
HVL	4	4	0	5	1	3	1	12	8	2.5
VLH	1	2	1	0	1	0	1	0	1	1
LVH	0	14	14	3	3	2	2	24	24	10.75

Speaker 10; File duration: 43 seconds

Table 42: Results of manual vs. system generated pitch marking for 4th file

	Automated	HE1	Error HE1	HE2	Error HE2	HE3	Error HE3	HE4	Error HE4	Average Error
F	34	31	3	12	22	0	34	28	6	16.25
HL	35	40	5	90	55	90	55	128	93	52
LH	34	30	4	38	4	44	10	78	44	15.5
VHL	1	13	12	1	0	2	1	3	2	3.75
HVL	4	1	3	11	7	2	2	7	3	3.75
VLH	6	0	6	0	6	0	6	0	6	6
LVH	0	14	14	2	2	2	2	20	20	9.5

It can clearly be observed from Tables 39-42 (Speaker 10) that the average error for F, HL and LH segment is very high as compared to the other four. HL segment has the maximum average error and VLH has the least average error except for the last file.

Next chapter includes the conclusions arrived at in this work and also some future indicators for carrying out such a work.

CHAPTER 5

Conclusion and Future Scope

5.1 Conclusion

Speech Recognition is a fascinating field spanning several areas of computer science and mathematics. The use of prosodic knowledge in automatic speech recognition is a widely researched topic in recent years. To incorporate prosody knowledge into automatic speech recognition by marking the variations in pitch is the main objective of this thesis. Pitch accent marking is an important parameter of sound signals. For pitch accent marking, voiced and unvoiced regions of speech are separated because pitch marking applies only to the voiced parts of the speech.

In this work for pitch accent marking, data has been collected in two different modes: read speech mode and lecture mode. A total of 40 files are taken from 10 different speakers: 5 male and 5 female members. Each speaker has contributed 4 files. Total duration of data considered in this work is 32 minute and 13 seconds. In read speech mode, data has been collected from seven different Punjabi speakers for a duration of about 22 minutes and in lecture mode, collected from three different Punjabi speakers for a duration of about 10 minutes. These files are saved in wave file format.

After collecting the data, pitch marking algorithm which has been implemented in MATLAB is applied to all the 40 files. In addition, 4 different human evaluators manually did the pitch marking of each file. In this work, 7 different labels are used to show the variations in pitch: F (Flat region), LH (Low to high region), HL (High to low region), VLH (Very low to high region), HVL (High to very low region), VHL (Very high to low region) and LVH (Low to very high region). The results are then presented in the form of tables.

5.2 Future Scope

Still improvements need to be done in the algorithm to achieve a high level of accuracy. It is miles away from perfection. Following are the improvements which need to be made.

1. The presented algorithm included F, LH, HL, VLH, HVL, VHL and LVH as labels for pitch marking. These parameters can be increased for higher accuracy. Additional parameters can be HVH (High to very high region), LVL (Low to very low region), VHH (Very high to high region), VLL (Very low to low region), etc.
2. An improved algorithm needs to be developed for proper division of the slots in the marking process as the algorithm presented in this thesis does not mark the slots accurately.
3. This system can also be extended for developing similar interface for other Indian languages.

REFERENCES

[1] Ananthapadmanabha, T., and Yegnanarayana, B. "Epoch Extraction of Voiced Speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 6, pp. 562-570, 1975.

[2] Atal, B. S. and Rabiner, L. R. "A pattern recognition approach to voiced-unvoiced-silence classification with applications to Speech Recognition," *IEEE transactions on Acoustic, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201-212, 1976.

[3] Rabiner, L., Wilpon, J.G., and Soong, F.K. "High performance connected digit recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 8, pp. 1214-1225, 1989.

[4] Brugnara, F., Falavigna, D., and Omologo, M. "Automatic segmentation and labelling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357-370, 1993.

[5] Ohmura, H. "Fine Pitch Contour Extraction by Voice Fundamental Wave Filtering Method," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 189-192, 1994.

[6] Fach, M., and Wokurek, W. "Pitch Accent Classification of Fundamental Frequency Contours by HMM," 1995.

[7] Harbeck, S., Kießling, A., Kompe, R., Niemann, H., and Nöth, E. "Robust pitch period detection using dynamic programming with an ANN cost function," *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 2, pp. 1337-1340, 1995.

[8] Slobada, T., Waibel and Alex. "Dictionary learning for spontaneous speech recognition," *4th International Conference on Spoken Language (ICSLP)*, vol. 4, pp. 2328-2331, 1996.

[9] Goncharoff, V., and Gries, P. "An Algorithm for Accurately Marking Pitch Pulses in Speech Signals," *Proceedings of the Signal Processing System*, 1998.

[10] Mann, I., and McLaughlin, S. "A nonlinear algorithm for epoch marking in speech

signals using poincare maps," *9th European Signal Processing Conference (EUSIPCO)*, pp. 1-4, 1998.

[11] Laprie, Y., and Colotte, V. "Automatic pitch marking for speech transformations via TDPSOLA," *9th European Signal Processing Conference (EUSIPCO), Rhodes*, pp. 1-4, 1998.

[12] Dikshit, P. "An Algorithm for locating fundamental frequency markers in Speech Signals," M.S. Thesis, Mangalore University, 2000.

[13] Colotte, V., and Laprie, Y. "Automatic enhancement of speech intelligibility," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1057-1060, 2000.

[14] Sakamoto, M., and Saito, T. "An Automatic Pitch-Marking Method using Wavelet Transform," *Proceedings of International Conference on Spoken Language Processing (ICSLP), China*, vol. 3, pp. 650-653, 2000.

[15] Veldhuis, R. "Consistent pitch marking," *International Conference on Speech language Processing*, vol. 3, pp. 207-210, 2000.

[16] Colotte, V., and Laprie, Y. "Higher precision pitch marking for TD-PSOLA," *11th European Signal Processing Conference*, pp. 1-4, 2002.

[17] Xijun, M., Zhang, W., Shi, Q., Zhu, W., and Shen, L. "Automatic prosody labelling using both text and acoustic information," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 516-519, 2003.

[18] Dibazar, A.A., Dong Song, Yamada, W., and Berger, T.W. "Speech Recognition Based on Fundamental Functional Principles of the Brain," *IEEE International Joint Conference on Neural Networks*, vol. 4, pp. 3071-3075, 2004.

[19] Ashouri, K., and Savoji, M.H. "Automatic and accurate pitch marking of speech signal using an expert system based on logical combinations of different algorithms outputs," *12th European Signal Processing Conference*, pp. 995-998, 2004.

[20] Ananthakrishnan, S., and Narayanan, S.S. "An Automatic Prosody Recognizer using a Coupled Multi-Stream Acoustic Model and a Syntactic-Prosodic Language Model," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 269-272, 2005.

[21] Dikshit, P., Zahorian, S.A., and Nagulapati, S. “An Algorithm for locating fundamental frequency markers in Speech Signals,” *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP),* pp. 33-36, 2005.

[22] Sjolander, K., and Beskow, J. (2006). *WaveSurfer User Manual* [Online]. Available: http://www.speech.kth.se/WaveSurfer/man.html

[23] Deng, H. and Shaughnessy, O. “Voiced-Unvoiced-Silence speech sound classification based on unsupervised learning,” *IEEE International Conference on Multimedia and Expo,* pp. 176-179, 2007.

[24] Chiang, C., Wang, X.D., Liao, Y., Wang, Y., Sin-Horng, C., and Hirose, K. “Latent Prosody Model of Continuous Mandarin Speech,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pp. 625-628, 2007.

[25] Rapp, B. “N-gram language models for Polish language. Basic concepts and applications in automatic speech recognition systems,” *Proceedings of the International Multiconference on Computer Science and Information Technology*, pp. 321–324, July 2008.

[26] Ananthakrishnan, S., and Narayanan, S. “A novel algorithm for unsupervised prosodic language model adaptation,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4181-4184, 2008.

[27] Chonjia, N., Liu, W., and Xu, B. “Automatic Prosody Boundary Labeling of Mandarin Using Both Text and Acoustic Information,” *6th International Symposium on Chinese Spoken Language Processing,* pp. 1-4, 2008.

[28] Sri Rama Murty, K., and Yegnanarayana, B. “Characterization of Glottal Activity from Speech Signals,” *IEEE Signal Processing Letters*, vol. 16, no. 6, 2009.

[29] Chalamandaris, A., Tsiakoulis, P., Karabetsos, S., and Raptis, S. “An efficient and robust pitch marking algorithm on the speech waveform for TD-PSOLA*,” IEEE International Conference on Signal and Image Processing Applications (ICSIPA),* pp. 397-401, 2009.

[30] Hayes, B. (2010). *Some Hints on Using WaveSurfer* [Online]. Available: http://www.linguistics.ucla.edu/people/hayes/103/WaveSurferHints

[31] Kivumbi. (2010). *Difference between Tone and Pitch* [Online]. Available:

http://www.differencebetween.net/miscellaneous/difference-between-tone-and-pitch

[32] Sheikhan, M., Safdarkhani, M.K., and Gharavian, D. "Presenting and classification based on three basic speech properties, using Haar wavelet analyzing," *2nd International Conference on Signal Processing Systems (ICSPS)*, vol. 3, pp. 189-191, 2010.

[33] Qian, Y., Wu, Z., Ma, X., Soong, F. "Automatic prosody prediction and detection with Conditional Random Field (CRF) models," *7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 135-138, 2010.

[34] Alias, F., and Munné, N. "Reliable Pitch Marking of Affective Speech at Peaks or Valleys Using Restricted Dynamic Programming," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 481-489, 2010.

[35] Sakshat Virtual Labs. (2011). *Identification of Voiced/Unvoiced/Silence Regions of Speech* [Online].
Available: http://iitg.vlab.co.in/?sub=59&brch=164&sim=613&cnt=1

[36] Mitra, V., Hosung Nam, Espy-Wilson, C.Y., Saltzman, E., and Goldstein, L. "Articulatory Information for Noise Robust Speech Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 1913-1924, 2011.

[37] Wang, M., Li, Y., Lin, M., Li, A., and Xiong, L. "The development of a database of functional and emotional intonation in Chinese," *International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pp. 136-141, 2011.

[38] Sharma, P. "Automatic Identification of Silence, Voiced and Unvoiced Chunks in Speech," M.S. thesis, Thapar University, Patiala, Punjab, 2012.

[39] Dua, M., Aggarwal, R.K., Kadyan, V., and Dua, S. "Punjabi Automatic Speech Recognition Using HTK," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 4, July 2012.

[40] Sin-Horng, C., Yang, J.H., Chiang, C., Liu, M., and Wang, Y. "A New Prosody-Assisted Mandarin ASR System," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1669-1684, 2012.

[41] Ning, M., Hongzhi, Y., Zhang, J., and Fang, H. "Prosody analysis on Tibetan Lhasa dialect news reading," *IEEE International Conference on Computer Science*

and Automation Engineering (CSAE), vol. 1, pp. 288-291, 2012.

[42] Brognaux, S., Drugman, T., and Beaufort, R. "Automatic detection and correction of syntax-based prosody annotation errors," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 410-415, 2012.

[43] Chang, L., Xu, J., Tang, K., and Cui, H. "A new robust pitch determination algorithm for telephone speech," *International Symposium on Information Theory and its Applications (ISITA*), pp. 789-791, 2012.

[44] Izzad, M., Jamil, N., Bakar, and Abu, Z. "Speech/non Speech detection in Malay language spontaneous speech," *International Conference on Computing, Management and Telecommunications*, pp. 219-224, 2013.

[45] Ykhlef, F., and Bendaouia, L. "Pitch Marking Using the Fundamental Signal for Speech Modifications via TD-PSOLA," *IEEE International Symposium on Multimedia (ISM)*, pp. 118-124, 2013.

[46] Sreejith, A., Mary, L., Riyas, K.S., Joseph, A., and Augustine, A. "Automatic prosodic labelling and broad class Phonetic Engine for Malayalam*," International Conference on Control Communication and Commuting (ICCC)*, pp. 522-526, 2013.

[47] Buza, O., Toderean, G., Balogh, A., and Domokos, J. "Algorithm for detection of voice signal periodicity," *7th Conference on Speech Technology and Human – Computer Dialogue (SpeD)*, pp. 1-5, 2013.

[48] Mittal, S. "Development of Phonetic Engine for Punjabi Language," M.S. thesis, Thapar University, Patiala, Punjab, 2014.

[49] Mankala, S. R., Bojja, S. R., Ramaiah, V. S., and Rao, R. R. "Automatic Speech Processing Using HTK for Telugu Language," *International Journal of Advances in Engineering and Technology*, vol. 6, no.6, pp. 2572-2578, 2014.

[50] Tamura, S., Seko, T., and Hayamizu, S. "Data Collection for Mobile Audio-visual Speech Recognition in Various Environments ," *17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA*), pp. 1-6, 2014.

[51] (2015). *Digital Audio Fundamentals* [Online]. Available: http://manual.audacityteam.org/o/man/digital_audio.html

[52] Lumenvox LLC. (2015). *Types of Speech Recognition* [Online]. Available: http://www.lumenvox.com/resources/tips/types-of-speechrecognition.aspx

[53] Wikipedia. (2015). *WaveSurfer* [Online]. Available: https://en.wikipedia.org/wiki/WaveSurfer

[54] Deekshitha, G., and Mary, L. "Prosodically guided phonetic engine," *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1-5, 2015.

VIDEO PRESENTATION:

https://www.youtube.com/watch?v=JUOEUF44XzU

APPENDIX

Snapshots of the GUI developed:

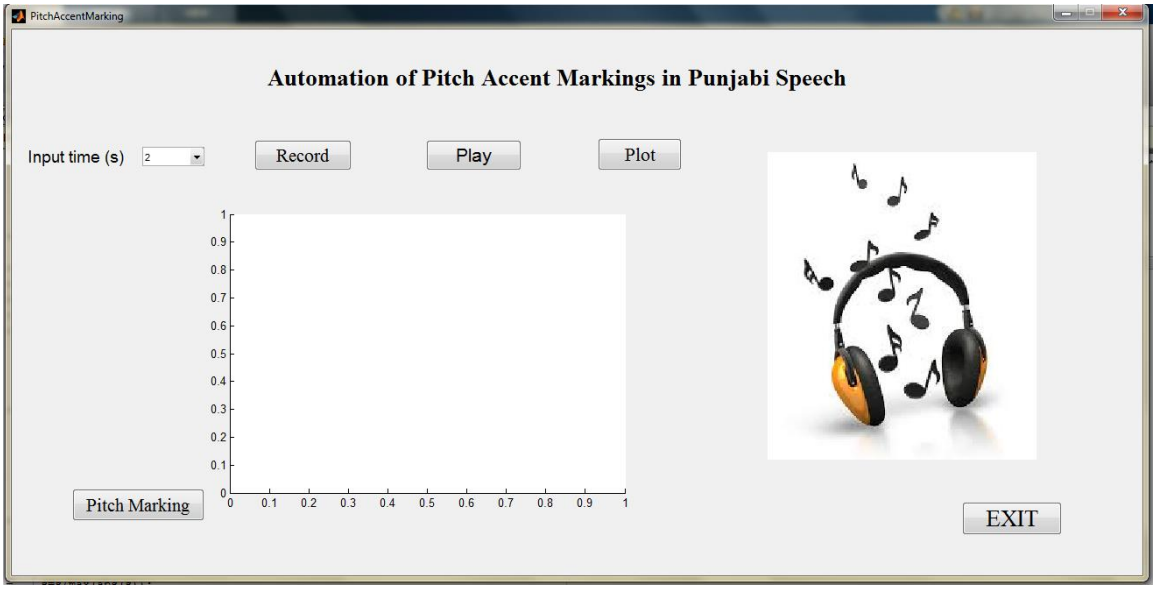

Figure 14: A view of GUI

Figure 15: Final output on WaveSurfer