An Acoustic Study of Tone Contrasts in Manipuri Language

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

Doctor of Philosophy

in

Computer Science and Engineering

by

Thiyam Susma Devi

Under the supervision of

Prof. Pradip K. Das

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

July, 2024

To

My Beloved Grandparents,
Bobok & Pupu

and

My Cherished Family

Acknowledgments

Completing my dissertation would have been insurmountable without the unwavering support and encouragement of numerous individuals for whom I am deeply grateful.

First and foremost, I extend my most profound gratitude to my esteemed supervisor, Professor Pradip K. Das, for his persistent support, patience and invaluable guidance throughout my doctoral research journey. His constant encouragement and positive mentorship have been instrumental in shaping me both as a researcher and an individual. I am truly indebted to him for the opportunity to work under his guidance, considering myself incredibly fortunate to have had this privilege. Professor Das has been more than a supervisor to me; he has been a source of inspiration, support and encouragement.

I express my heartfelt appreciation to the members of my doctoral committee, Professor Shivashankar B. Nair, Professor Konjengbam Darunkumar Singh, Dr. Rashmi Dutta Baruah and Dr. Pinaki Mitra, for their insightful feedback and constructive suggestions, which have significantly enriched the quality of my work.

I sincerely thank Professor Khumanthem Manglem Singh, Director in-charge of NIT Manipur, for his support in pursuing my Ph.D. at IIT Guwahati. I also want to acknowledge Professor Tamarapalli Venkatesh, Head of the Department at IITG, for providing me with the necessary facilities for my research.

Special acknowledgment goes to Prof. Chungkham Yashawanta Singh, Dr. Yumnam Aboy Singh, Waikhom Ibeyaima Devi, Thockchom Birjit Singha, Yengkhom Omesh Singh and Jennil Thiyam for their contributions to the development of the ManiTo speech corpus, as well as to all the speakers who participated in the recording process.

I am grateful to Mrs. Moon Das, Dr. Priya Nair, Prof. Hemangee K. Kapoor and Chongtham Gilbert for their invaluable advice and continuous encouragement, which have been a strength throughout my academic pursuit.

My heartfelt appreciation extends to all my colleagues and friends who have accompanied me throughout my Ph.D. journey, with special mention to my fellow lab mates at the Speech Lab, especially Dr. Bhagath Parabattina for his unwavering support and my friend Vanshali Sharma, who has been there for me in times of difficulty.

The facilities provided by the institute greatly enhanced the comfort of staying at the IITG campus and I am thankful for the vibrant campus life, enriched by the presence of my seniors Sanasam Vipez Devi, Nanaobi and friends Sonia, Vanshali, Karnish, Chitra, Komal, Ashita and others.

I am grateful for the support of my colleagues at NIT Manipur, particularly Sanasam Renubala Devi and Salam Devayani Devi, who have stood by me during challenging times.

Above all, it comes to my family, who have continuously cared for me and my well-being since childhood. I am deeply grateful to my grandmother, Soibam Manisang Devi, whose wisdom and love have profoundly shaped me. My grandfather, Yumnam

Mangoljao Singh, has provided invaluable guidance. My parents, Yumnam Memcha Devi and Thiyam Surchand Singh have unwaveringly supported my Ph.D. pursuit. My sisters, Surma Thiyam and Sonia Thiyam and my younger brother, Thiyam Meghabarna, have been constant sources of love and inspiration. I extend heartfelt appreciation to Pankaj Kumar Kalita for his joyful presence, unwavering support and continuous motivation, which have enriched my life.

Lastly, I am grateful to the Almighty God, whose guidance and blessings have illuminated my path.

July, 2024 Thiyam Susma Devi

Declaration

I certify that:

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor.
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisors are not in a position to check for any possible instance of plagiarism within this submitted work.

July, 2024 Thiyam Susma Devi

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Prof. Pradip K. Das
Professor
Email : pkdas@iitg.ac.in
Phone : +91 361 2582353

Certificate

This is to certify that this thesis entitled "**An Acoustic Study of Tone Contrasts in Manipuri Language**" submitted by **Thiyam Susma Devi**, in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy to the Indian Institute of Technology Guwahati, Assam, India, is a record of the bonafide research work carried out by her under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam, India. To the best of my knowledge, no part of the work reported in this thesis has been presented for the award of any degree at any other institution.

Date: July, 2024
Place: Guwahati

Prof. Pradip K. Das
(Thesis Supervisor)

An Acoustic Study of Tone Contrasts in Manipuri Language

Abstract

Speech is a natural and intuitive mode of human communication, underscoring the essence of interpersonal interaction. Automatic Speech Recognition (ASR) is a pivotal innovation in digital technology, empowering devices to comprehend and process spoken language seamlessly. ASR's applications span various domains, including dictation software, voice-activated assistants and automated call centers, thus revolutionizing how we engage with technology. Its significance extends further to the development of assistive devices for individuals with disabilities and the preservation of endangered languages, wherein ASR catalyzes documentation and linguistic conservation.

Manipuri is a low-resource Tibeto-Burman tonal language primarily spoken in the northeastern state of Manipur, India. Tone identification is crucial to speech comprehension for tonal languages, where tone defines the word's meaning. ASR for those languages can perform better by including tonal information from a powerful tone detection system. Despite extensive research on tonal languages such as Mandarin, Thai, Cantonese and Vietnamese, there is a significant gap in exploring Manipuri's tonal features.

This thesis presents the development of a meticulously crafted speech corpus called ManiTo, explicitly designed to analyze the tones of Manipuri. Comprising 17,837 labeled audio samples from twenty native speakers, ManiTo facilitates a nuanced examination of Manipuri's tonal contrasts. Initial investigations reveal the presence of two distinct tones: Falling and Level. A comprehensive acoustic feature analysis was conducted to differentiate between the two tones to deepen our understanding. Two sets of features, focusing on pitch contours, jitter and shimmer measurements, were explored to delineate Manipuri's tonal nuances. Various classification algorithms were employed to validate the selected feature sets, including Support Vector Machine, Long Short-Term Memory, Random Forest and k-Nearest Neighbors. Results demonstrate that the second feature set consistently outperformed the first, especially when utilizing the Random Forest classifier. These findings provide crucial insights for advancing speech recognition technology in low-resource tonal languages like Manipuri. This thesis contributes to the broader understanding of tonal languages through the development of ManiTo and the insights gained from acoustic feature analysis. It sets the stage for future research to enhance speech recognition technologies in linguistically diverse and underrepresented languages.

Contents

List of Figures

List of Tables

List of Abbreviations

Terms	Abbreviations
AMDF	Average Magnitude Difference Function
ANN	Artificial Neural Network
apq3	Three-point Amplitude Perturbation Quotient
apq5	Five-point Amplitude Perturbation Quotient
apq11	Eleven-point Amplitude Perturbation Quotient
ASR	Automatic Speech Recognition
BNF	Bottleneck Feature
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DBNFs	Deep Bottleneck Features
δF_0	Delta of Fundamental Frequency
DBN	Deep Belief Network
ddp	Difference of Differences of Periods
DNN	Deep Neural Networks
DWT	Discrete Wavelet Transforms
EER	Equal Error Rate
EPS	Entropy Based Phonotactic Segmentation
F_0	Fundamental Frequency
F_0_init	Initial Fundamental Frequency
F_0_final	Final Fundamental Frequency
F_0_max	Maximum value of Fundamental Frequency
F_0_min	Minimum value of Fundamental Frequency
F_0_ran	Fundamental Frequency range
FFSs	Feature Fusion Sets
GMM	Gaussian Mixture Model
GSR	Google Speech Recognition
HCRF	Hidden Conditional Random Field
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
IVR	Interactive Voice Response
KNN	k-Nearest Neighbors
KWS	Keyword Search
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coding
LPCs	Linear Prediction Coefficients
LPCC	Linear Predictive Cepstral Coefficients
LPCM	Linear Pulse Code Modulation
LSTM	Long Short-Term Memory
LVCSR	Large Vocabulary Continuous Speech Recognition
MFCC	Mel Frequency Cepstral Coefficient

MFPLP	Mel Frequency with Perceptual Linear Prediction
MLLT	Maximum Likelihood Linear Transformation
MLP	Multi-Layer Perceptron
MSD	Multi Space Distribution
MSVM	Multi-class Support Vector Machine
NCC	Normalized Cross-Correlation
NN	Neural Networks
PCA	Principal Component Analysis
PE	Phonetic Engine
PER	Phone Error Rates
PLP	Perceptual Linear Predictive
PNCC	Power-Normalized Cepstral Coefficients
POS	Part of Speech
ppq5	Five point Period Perturbation Quotient
PPRLM	Parallel Phone Recognition followed by Language Model
rap	Relative Average Perturbation
RBCOM	Rule Based Corpus Optimization Model
RF	Random Forest
RNN	Recurrent Neural Networks
SCSC	Syllable Corpus of Standard Chinese
SD	Standard Deviation
SGMM	Space Gaussian Mixture Mode
$sign\_F_0\_ran$	Sign of Fundamental Frequency range
$sign\_F_0\_slope$	Sign of Fundamental Frequency slope
SVM	Support Vector Machine
SY	Standard Yoruba
TDNN	Time Delay Neural Network
WER	Word Error Rate
VOV	Voice of Vietnam

1

Introduction

This chapter offers an overview of the thesis's content and layout while highlighting the key contributions discussed in the following chapters. Beginning with an introduction to the fundamental concepts of tone, we provide an insightful overview of Manipuri and its tone system. We then delve into the specific challenges posed by tone recognition in Manipuri, considering variations in tone patterns, contextual influences and speaker variability. Our motivation is driven by the need to address these challenges and contribute to advancing technology and linguistic understanding. In the final sections, we explore potential avenues to mitigate these challenges and summarize the research's contributions to tone recognition in Manipuri, aiming to facilitate technological progress and deepen our comprehension of tonal languages.

1.1 Tone

In linguistics, tone refers to pitch patterns to distinguish between words or convey grammatical information. Pitch variations can affect a word's meaning in tonal languages, unlike non-tonal languages, which could imply intonation or emotional emphasis.

Yip [93] defines tonal language as a language where an indication of pitch enters into the linguistic realization of at least some morphemes. Each syllable or word in a tonal language is associated with a specific pitch contour or tone, contributing to its lexical or grammatical meaning.

Tone can be associated with fundamental frequency (F_0) and pitch, the perceptual correlate of fundamental frequency. F_0 refers to the number of pulses per second contained within a signal, where each pulse corresponds to a single vocal cord vibration. This frequency is measured in Hertz (Hz), with one Hz representing one cycle per second. Pitch, however, represents the perceived F_0, determining whether a signal is heard as high or low in pitch. Therefore, pitch forms the foundational basis for understanding tone.

For instance, in Mandarin Chinese, there are four lexical tones: flat high, rising, falling-rising and falling. The word 'ma' with different tones can mean 'mother,' 'hemp,' 'horse' or 'scold' depending on the tone. Similarly, in languages like Thai, Vietnamese and Yoruba, among others, tones play a crucial role in distinguishing between words.

In essence, tone in tonal languages is a phonetic feature where variations in pitch serve as meaningful linguistic elements, contributing to the overall structure and comprehension of the language.

1.2 Manipuri

Manipuri is one of the scheduled Indian Tibeto-Burman languages spoken predominantly in Manipur, a northeastern state of India [8]. It is recognized as one of the 22 official languages of India [52, 15]. Beyond its primary stronghold, Manipuri is spoken by minority communities residing in two additional Indian states, namely Assam and Tripura. Moreover, its influence extends across borders, with small populations conversing in Manipuri in neighboring countries such as Bangladesh and Myanmar (Burma).

Meetei Mayek is the script of Manipuri. It consists of 27 main alphabets called Mapung Mayek, 8 unreleased characters called Lonsum Mayek, 8 vowel signs called

Fig. 1.1: Manipur in the India map with neighboring countries

Cheitap Mayek, 3 punctuation marks including diacritics and Cheising Mayek for the numerical. The phonological system of Manipur speech contains three significant types of sounds: consonants, vowels and tones [74]. The balance of the phonological structure has to do with the inter-relationships of these elements and systems and how they are combined to form syllables and pause groups. The sound system of Manipuri consists of twenty-four consonants, twelve vowels (six monophthongs and six diphthongs) and two tones [13, 74, 76, 80]. The two types of tone in Manipuri are:

i) **Falling** tone

ii) **Level** tone

Every syllable bears one of these two tones. In the Manipuri transcript, the falling tone is noted by the tone marker, lum mayek (·), just after the syllable and with the grave accent (`) symbol above the vowel in English representation, while the level tone is unmarked. An example of a Manipuri tone contrast word pair is shown in Fig. 1.2 and Fig. 1.3.

Figure 1.4 illustrates the pitch contour of the two tones in Manipuri. The orange line represents the falling tone, whereas the level tone is depicted in blue. The falling tone

Falling tone

Fig. 1.2: ꯁꯤꯡ/sìŋ/ meaning firewood

Level tone

Fig. 1.3: ꯁꯤꯡ/siŋ/ meaning ginger

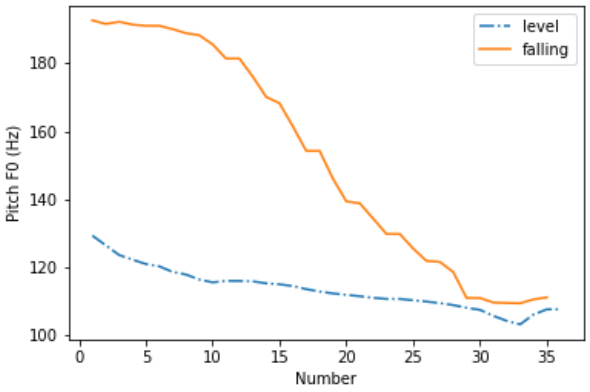


Fig. 1.4: Pitch contour of two tones in Manipuri

initiates at a high frequency and gradually decreases, whereas the pitch of the level tone remains relatively constant.

1.3 Motivation

Progressive research has been done on tonal languages such as Mandarin, Chinese, Thai and Vietnamese and some work has been reported in Bodo, Mizo, Punjabi, Lithuanian, Latvian, Hausa and Yoruba [73], but Manipuri needs to be explored. For tonal languages, tone gives an implicit character to a word. So, a successful tone recognition system can significantly improve Manipuri's speech understanding tasks. Manipur is a hub for tourism and trade in the northeastern region of India. A speech recognition system for Manipuri can help boost the local economy by improving communication between

tourists and locals, facilitating trade and commerce and empowering people by giving them a tool to communicate and access information more effectively. It can improve their quality of life and enhance their sense of agency and control. Manipuri is the sole language of communication in Manipur among the 29 various ethnic communities [34]. Developing a speech recognition system for Manipuri can help facilitate communication and improve accessibility for the people living in Manipur. Automatic Speech Recognition (ASR) has the potential to improve accessibility, increase productivity, enhance user experience and facilitate communication across a wide range of contexts and applications. Integration of tone recognition in ASR systems is vital for unlocking the full potential of tonal languages in the digital landscape. This motivates us to study the tone characteristics of Manipuri to develop a tone detection system for a robust ASR for the Manipuri language.

1.4 Challenges in Tone Recognition for Manipuri

Tone recognition in Manipuri presents unique challenges due to low-resources, variations in tone patterns, contextual influences and speaker variability.

- **Limited Dataset Availability:** One of the primary challenges in tone recognition for Manipuri is the lack of comprehensive datasets. Developing robust tone recognition systems becomes challenging without a diverse and well-annotated dataset capturing the full range of Manipuri tonal patterns and variations. Limited data availability hampers the training and evaluation of models, hindering progress in tone recognition research.

- **Complex Tone Patterns**: Manipuri is a tonal language with intricate tone patterns, where variations in pitch contour play a crucial role in distinguishing lexical and grammatical meanings. Without sufficient data and prior research, accurately

modeling these intricate tonal patterns remains a significant challenge.

- **Contextual Influences**: Various linguistic contexts can influence Manipuri tones, such as co-articulation between syllables, sentence structure and syntactic patterns. Without robust contextual modeling, tone recognition systems may misinterpret tones or fail to capture subtle tonal variations, particularly in connected speech.

- **Speaker Variability**: Speaker variability encompasses many factors, including accent, pitch range and speaking style differences. In tonal languages like Manipuri where pitch variations convey lexical and grammatical distinctions, gender differences in pitch become particularly relevant. Typically, males and females may exhibit distinct pitch ranges, with males often having a lower pitch than females. Consequently, gender-related differences in pitch can influence the realization of tones in speech, affecting the accuracy of tone recognition systems. Without accounting for gender differences in pitch and incorporating gender-balanced datasets, tone recognition models may struggle to generalize across diverse speaker populations, reducing performance and accuracy in real-world applications.

Without a comprehensive dataset and prior work on Manipuri tone recognition, developing tone recognition systems faces challenges in accurately capturing and modeling the diverse tonal patterns and variations in Manipuri speech. Without sufficient data and prior work to inform model development, addressing these challenges requires concerted efforts to collect high-quality datasets, conduct linguistic analyses and develop innovative methodologies tailored to the complexities of Manipuri tone recognition.

1.5 Goals and Contributions

Tonal languages, in particular, pose unique challenges and opportunities for computational linguistics, especially when confronted with limited linguistic resources. Low-resource environments often need comprehensive linguistic datasets and well-established infrastructures for training machine learning models. Recognizing tone in such languages is paramount despite the inherent difficulties, as it directly impacts communication, language preservation and performance of ASR systems. The following is the outline of the goals and contributions of our study:

- **Develop an isolated speech dataset consisting of tonal contrast words in Manipuri, a low-resource tonal language:** We have initiated the creation of a hand-crafted speech corpus focusing on tonal contrasts in Manipuri. This corpus was developed from recordings of 20 native speakers, resulting in 17,837 samples. Further, two variants of ManiTo are developed: ManiToSpc, which features spectrograms derived from ManiTo's WAV files, providing a two-dimensional audio representation and ManiToScl, containing scalograms generated from the audio files. These variants offer potential applications for deep learning techniques that leverage image-based approaches in the analysis of audio data. Our efforts aim to address the limited availability of Manipuri speech data for tonal studies, providing valuable resources for research and analysis in this linguistic domain.

- **Analysis of Manipuri tones in the tonal contrast database, ManiTo:** The initial Manipuri Tonal Contrast dataset analysis suggests pitch contour effectively distinguishes between falling and level tones, with slope and height key factors. Additionally, gender-based analysis reveals non-comparability between male and female speech data, necessitating separate gender-based analysis.

- **Acoustic features analysis to disambiguate the tones of Manipuri:** We have proposed two feature sets: the first is derived from pitch contours and the second augments the first with jitter and shimmer features.

- **Modeling tone recognition system for Manipuri:** We have adopted four classifiers, Support Vector Machine, Long Short-Term Memory, k-Nearest Neighbors and Random Forest to evaluate their effectiveness of Manipuri tone recognition in gender-dependent and speaker-dependent scenarios. It is found that the pitch variation between the gender and the variation induced by co-articulation between syllables influence the tone recognition rate of the Manipuri tones. Comparative experiments demonstrate that when integrated with the second feature set, the Random Forest classifier exhibits superior performance in high accuracy, simplicity and robust learning capabilities for the Manipuri tone recognition system.

This study aims to propel Manipuri tone research forward by constructing a comprehensive speech corpus and determining optimal feature sets to evaluate tone recognition model performance. This endeavor addresses a critical gap in current research on Manipuri tone, laying the groundwork for enhanced understanding and analysis.

1.6 Thesis Layout

The thesis consists of seven chapters organized as follows:

- **Chapter 1** introduces the thesis's content layout and emphasizes its key contributions.

- **Chapter 2** provides background information on fundamental concepts essential for later discussions in the thesis.

- **Chapter 3** offers an overview of related works on tone recognition.

- **Chapter 4** details the development of the ManiTo speech corpus.
- **Chapter 5** performs the preliminary analysis of the Manipuri tones in the tonal contrast dataset, ManiTo.
- **Chapter 6** investigates features and models for tone detection, experimental analysis and results.
- **Chapter 7** concludes the thesis by summarizing the findings and discussing future research directions.

2

Background

Speech is one of the most natural and convenient means of communication, allowing individuals to convey thoughts, emotions and intentions effortlessly. With its roots deeply embedded in human history, speech is a fundamental mode of interaction across cultures and societies. From everyday conversations to public speeches and professional presentations, the power of spoken language permeates every aspect of human interaction, facilitating the exchange of ideas and the building of relationships.

2.1 Automatic Speech Recognition

Technological advancements have revolutionized how we interact with speech in recent years, paving the way for the development of Automatic Speech Recognition (ASR) systems. ASR systems enable the conversion of spoken language into text or commands, offering unprecedented convenience and accessibility in various domains. By harnessing the capabilities of machine learning, natural language processing and signal processing, ASR systems can accurately transcribe spoken words, even in noisy environments or with diverse accents.

2.1.1 Applications of ASR

Automatic Speech Recognition has applications in various domains, facilitating numerous tasks and enhancing user experiences. Here are some key applications of ASR:

- **Virtual Assistants:** ASR powers virtual assistants like Siri, Google Assistant and Amazon Alexa, enabling users to perform hands-free tasks such as setting reminders, sending messages, making calls and controlling smart home devices through voice commands.

- **Transcription Services:** ASR technology transcribes spoken audio into text format and facilitates tasks such as meeting minutes, lecture notes, interview transcripts and legal documentation. Transcription services powered by ASR enhance efficiency and accuracy in data processing and documentation.

- **Customer Service and Call Centers:** ASR is employed in customer service applications to automate call routing, provide Interactive Voice Response (IVR) systems and enable voice-based inquiries and transactions. ASR-driven call centers streamline customer interactions, reduce wait times and improve service quality.

- **Language Learning and Accessibility:** ASR technology supports language learning by providing pronunciation feedback, vocabulary drills and language exercises based on spoken input. Additionally, ASR enhances accessibility for individuals with disabilities by enabling voice commands for operating computers, smartphones and other assistive devices.

- **Dictation and Voice-to-Text:** ASR enables hands-free dictation and voice-to-text conversion, allowing users to dictate emails, documents and notes into text-based applications directly. Voice-to-text functionality enhances productivity and facilitates multitasking in various professional and personal contexts.

- **Navigation and Directions:** ASR is integrated into navigation systems and mapping applications to enable voice-guided directions and real-time traffic updates.

Voice-activated navigation enhances driver safety and convenience by minimizing distractions and providing hands-free interaction.

- **Voice Search and Information Retrieval:** ASR powers voice search functionality in search engines, mobile apps and smart devices, allowing users to retrieve information and perform online queries using spoken queries. Voice search enhances user experiences by providing fast and convenient access to relevant information.

- **Entertainment and Media:** ASR technology is utilized in entertainment and media applications for voice-controlled media playback, interactive storytelling and voice-activated content discovery. Voice-enabled entertainment systems offer immersive and personalized experiences for users across various platforms.

- **Healthcare and Medical Transcription:** ASR systems are employed in healthcare for medical dictation, transcription and documentation of patient records, procedures and clinical notes. ASR-driven medical transcription services streamline administrative tasks, improve documentation accuracy and enhance patient care coordination.

- **Security and Authentication:** ASR technology is utilized for voice-based authentication and identity verification in security systems, access control and biometric authentication applications. Voice-based authentication enhances security by providing an additional layer of identity verification based on unique vocal characteristics.

- **Education and accessibility tools for individuals with disabilities:** Education and accessibility tools powered by ASR technology enhance inclusively in learning environments. These tools enable individuals with disabilities to access educational content through speech-to-text transcription and real-time captioning

services. By integrating ASR, educational institutions promote equal access to learning opportunities and empower diverse learners to engage actively in the educational process.

Overall, ASR technology continues to drive innovation across diverse industries, revolutionizing how we interact with technology, access information and perform everyday tasks through the power of spoken language.

2.1.2 Architecture of ASR

The typical architecture of an ASR system [50] is shown in Fig. 2.1, which comprises four main modules: preprocessing, feature extraction, classification and language model.

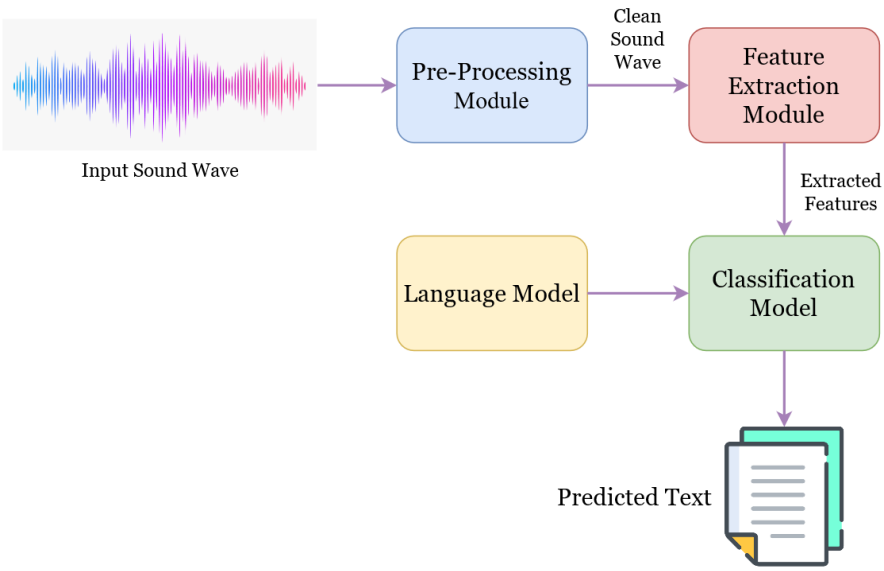


Fig. 2.1: A typical architecture of an ASR system

- **Preprocessing:** The preprocessing module aims to reduce noise in the input signal. Various methods such as segmentation, framing, endpoint detection and normalization are applied to the input audio to prepare it for feature extraction.

- **Feature Extraction:** Preprocessed speech undergoes feature extraction, where the extracted features significantly influence the efficiency and performance of the classification module. Commonly used methods include Discrete Wavelet Transform (DWT), Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC), etc.

- **Classification:** The classification module predicts the transcript of the input speech. It employs three classification methods: acoustic-phonetic, pattern recognition and artificial intelligence approaches. The first method is the acoustic-phonetic approach, which forms the joint probability distribution using the training inputs and uses it to predict the output. Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) are commonly used. The second approach is pattern recognition. It involves pattern training and pattern comparison. Dynamic Time Warping is one of the pattern recognition algorithms. The third is the artificial intelligence approach, which calculates the parametric model using the input dataset and their respective output vectors. Typical examples of this approach include Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

- **Language Model:** The language model module recognizes the phonemes predicted by the classifier. It encompasses various semantics and language rules. While ASR systems can function without language models, their performance improves significantly with their integration [10].

These modules work together to process and analyze spoken language, enabling accurate transcription and interpretation of speech in ASR systems.

2.1.3 Tone Recognition

Tone recognition refers to identifying and categorizing the pitch patterns and variations in spoken language, particularly in tonal languages where pitch differences convey lexical or grammatical distinctions. Tone recognition in ASR systems is critical in accurately transcribing and interpreting spoken language, particularly in tonal languages such as Mandarin Chinese, Thai and Manipuri, where pitch variations convey a word's meaning. Incorporating tone information into ASR models enables them to capture the nuances of tonal languages, facilitating more precise transcription and interpretation of speech. By analyzing pitch contours and patterns, ASR systems can identify and distinguish different tones, enhancing the accuracy and effectiveness of speech recognition. The goal of tone recognition is to accurately interpret the tonal characteristics of spoken language and correctly identify the intended meaning conveyed by the speaker. Tone recognition has applications in various fields, including ASR, language learning, linguistic research and human-computer interaction. As ASR technology evolves, advancements in tone recognition algorithms and techniques contribute to more robust and reliable speech recognition systems, further enhancing their usability and applicability across various domains.

2.2 Manipuri

Manipuri (Meiteiron) stands out as India's sole scheduled Tibeto-Burman language, distinguished by its unique script known as Meitei/Meetei Mayek, commonly called the 'Meitei script.' Unlike others, this language enjoys the privilege of having its indigenous writing system. It enjoys widespread usage in the northeastern Indian state of Manipur, with smaller communities also conversing in Manipuri in the states of Assam and Tripura, as well as in neighboring countries like Bangladesh and Myanmar. Spoken by

approximately 1.5 million people [8], Manipuri is traditionally written using Bengali and Meitei script. The Meitei script, utilized for writing the Manipuri language from 33 AD until the early 18th century [54], underwent a significant shift in 1717. Under the directive of Manipur's then-ruler, the Meitei script was supplanted by the Bengali script for official Manipuri documents. Despite efforts by researchers to reintroduce a modernized version of the ancient Meitei script in 1980, it was not until 2009 that the script was officially encoded in Unicode. Since then, the Meitei script has gradually gained prominence in various domains, including education, administration and commerce, alongside the Bengali script. The original Meitei script is anticipated to reclaim its position as the primary script for composing Manipuri documents in the coming years.

2.2.1 Manipuri Script

The indigenous script of Manipuri, Meitei Mayek, comprises 27 main letters, commonly known as Mapung Mayek, 18 of which are indigenous and 9 exotic. Additionally, there are 8 Lonsum Mayek, referred to as unreleased letters, 8 Cheitap Mayek encompassing vowel signs including matra and 3 Khudam Mayek, incorporating punctuation marks including diacritics. Notably, Manipuri features a numerical system called Cheising Mayek [76, 1]. Table 2.1 illustrates the Mapung Mayek with its phonetic representation, while Table 2.2 displays the Lonsum Mayek. In Table 2.3, the Cheitap Mayek is presented, followed by the Khudam Mayek in Table 2.4 and finally, Table 2.5 exhibits the Cheising Mayek.

2.2.2 Phonological System of Manipuri

Phonology is the study of sound patterns that occur within languages [7]. In phonology and linguistics, a phoneme is a distinct sound in a specified language that distinguishes one word from another [88]. Phonological features can be used effectively in an ASR

Table 2.1: Mapung Mayek (main letters)

Sl. No.	Script	Name	Phonetic Representation
1	ꯀ	kok	k
2	ꯁ	sam	s
3	ꯂ	lai	l
4	ꯃ	mit	m
5	ꯄ	paa	p
6	ꯅ	naa	n
7	ꯆ	chil	c
8	ꯇ	til	t
9	ꯈ	khou	k^h
10	ꯊ	ngou	ŋ
11	ꯊ	thou	t^h
12	ꯋ	wai	w
13	ꯌ	yang	y
14	ꯍ	huk	h
15	ꯎ	un	u
16	ꯏ	i	i
17	ꯐ	pham	p^h
18	ꯑ	atiya	a
19	ꯒ	gok	g
20	ꯖ	jham	j^h
21	ꯔ	rai	r
22	ꯕ	baa	b
23	ꯖ	jil	j
24	ꯗ	dil	d
25	ꯘ	ghou	g^h
26	ꯗ	dhou	d^h
27	ꯙ	bham	b^h

Table 2.2: Lonsum Mayek (unreleased characters)

Sl. No.	Script	Name	Phonetic Representation
1	ꯀ	kok lonsum	k
2	ꯜ	lai lonsum	l
3	ꯃ	mit lonsum	m
4	ꯄ	pa lonsum	p
5	ꯅ	na lonsum	n
6	ꯇ	til lonsum	t
7	ꯉ	ngou lonsum	ŋ
8	ꯏ	i lonsum	i

Table 2.3: Cheitap Mayek (vowel signs including matra)

Sl. No.	Script	Name	Phonetic Representation
1	ꯀꯥ꯫	aatap	a
2	ꯀꯤ꯫	inap	i
3	ꯀꯨ꯫	uunap	u
4	ꯀꯦ꯫	yetnap	e
5	ꯀꯩ꯫	cheinap	əi
6	ꯀꯣ꯫	otnap	o
7	ꯀꯧ꯫	sounap	əu
8	ꯀꯪ꯫	nung	ŋ

Table 2.4: Khudam Mayek (Punctuation Marks including Diacritics)

Sl. No.	Script	Name	Use
1	꯫	cheikhei	full stop
2	꯭	apun	cluster mark
3	꯰	lum	tone marker

Table 2.5: Cheising Mayek (numerals)

Sl. No.	Script	Name	International numeral
1	ꯐ	phul	0
2	ꯑ	ama	1
3	ꯍ	ani	2
4	ꯎ	ahum	3
5	ꯃ	mari	4
6	ꯄ	manga	5
7	ꯇ	taruk	6
8	ꯊ	taret	7
9	ꯅ	nipan	8
10	ꯆ	mapan	9

system for pronunciation training [3]. A phonological feature-based ASR system allows the authors to perform a sub-phonemic analysis at the feature level, providing more effective feedback to reach the acoustic goal and perceptual constancy. It constrains what can be expected in the speech stream [87]. The phonological system of Manipuri

Table 2.6: The place of articulation in Manipuri consonants.

Manner	Places								
	Bilabial		Alveolar		Palatal		Velar		Glottal
Plosive	p	b	t	d	c	ɟ	k	g	
Plosive Aspirated	p^h	b^h	t^h	d^h		$ɟ^h$	k^h	g^h	
Nasals		m		n				ŋ	
Fricative				s					h
Approximant		w		ɹ		j			
Lateral Approximant				l					

speech encompasses 38 phonemes, consisting of 36 segmental and 2 suprasegmental phonemes. Of these 38 phonemes, 24 are consonants, 6 are monophthongs, 6 are

diphthongs and 2 are tones [76, 74, 75]. The segmental phonemes include consonants, monophthongs and diphthongs, while the suprasegmental phonemes pertain to tones. This comprehensive phonemic inventory forms the foundation of Manipuri language structure and communication.

- **Consonant:** The Manipuri consonant inventory consists of twenty-four consonants. Table 2.6 shows the place of articulation in Manipuri consonants.

- **Vowel:** There are six vowels in Manipuri.

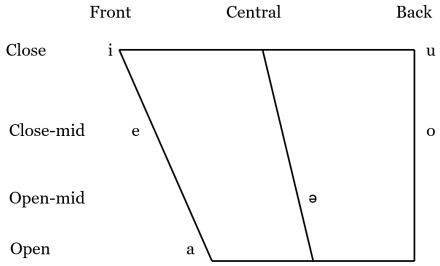


Fig. 2.2: Monophthongs chart

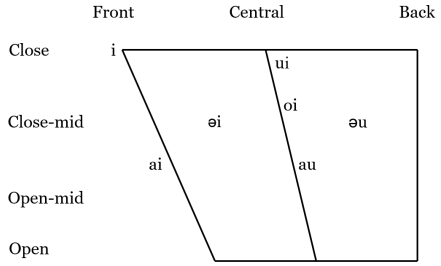


Fig. 2.3: Diphthongs chart

  - Monophthongs: Manipuri features six monophthongs representing single vowel sounds. Their mapping to the International Phonetic Alphabet (IPA) chart is illustrated in Fig. 2.2.
  - Diphthongs: Additionally, there are six diphthongs in Manipuri, where two single vowel sounds are combined. The mapping of these diphthongs to the IPA chart is depicted in Fig. 2.3.

- **Tone:** Manipuri is classified as a tonal language, with a lexically significant and contrastive pitch on each syllable. It distinguishes between two tones:
  - A level tone, which is unmarked in the transcript.
  - A falling tone, identified by the lum mayek diacritic symbol "•".

Table 2.7: A list of few Manipuri tonal contrast word pairs and their meanings

Sl. No.	Falling Tone Word	Meaning	Level Tone Word	Meaning
1	ꯎꯟ꯭ꯅ /ùn/	skin	ꯎꯟꯅ /un/	ice, snow
2	ꯊꯣꯡ꯭ /tʰòŋ/	door	ꯊꯣꯡ /tʰoŋ/	bridge
3	ꯄꯨ꯭ꯕ /pùbə/	to borrow	ꯄꯨꯕ /pubə/	to bring
4	ꯆꯤ꯭ /cà/	wax / birthmark	ꯆꯤ /ca/	tea
5	ꯈꯣꯏ꯭ /kʰòi/	navel	ꯈꯣꯏ /kʰoi/	bee / fishing hook
6	ꯁꯤꯡ꯭ /sìŋ/	firewood	ꯁꯤꯡ /siŋ/	ginger
7	ꯆꯤ꯭ꯕ /càbə/	eating	ꯆꯤꯕ /cabə/	swim visibly
8	ꯊꯨ꯭ꯕ /lùbə/	clear	ꯊꯨꯕ /lubə/	difficult
9	ꯊꯨꯝ꯭ꯕ /lùmbə/	heavy	ꯊꯨꯝꯕ /lumbə/	warm
10	ꯄꯥꯏ꯭ꯕ /paìbə/	touch	ꯄꯥꯏꯕ /paibə/	fly

Table 2.7 shows some tone contrast word pairs of Manipuri.

In Manipuri, the pitch contour of each syllable holds lexically significant and contrastive properties, contributing to the overall semantics of the language. Unlike non-tonal languages, where pitch variations may serve primarily as intonation markers or grammatical elements, in Manipuri, subtle tone changes can completely change a word's meaning.

For instance, consider the word 'un' in Manipuri, which can have different meanings based on its tone. When pronounced with a falling tone, /ùn/ signifies 'skin,'

whereas a level tone transforms it into ‘snow.’ The tonal nuances in Manipuri thus necessitate precision in pronunciation and acute sensitivity to pitch variations during communication.

Understanding the tonal intricacies of Manipuri is not only essential for native speakers but also for learners and linguists interested in the language’s phonological structure and cognitive processes. Research into Manipuri tone recognition and production sheds light on the broader study of tonal languages, contributing valuable insights into human language perception, processing, communication strategies and developing a robust ASR system of Manipuri.

3

Literature Survey

Exploring natural language and human-machine interaction presents a widespread and intricate research avenue. The primary aim is to develop systems capable of organized communication with humans across various operational environments. Tone recognition in tonal languages has been a subject of extensive research driven by these languages' linguistic intricacies and computational challenges. Tonal languages are characterized by pitch variations, where pitch changes can alter a word's meaning. Understanding and accurately recognizing tones are crucial for various applications, including speech recognition systems, language preservation efforts and linguistic studies. Tonal languages can be categorized based on their language family. Table 3.1 provides an overview of the characteristics of the phonological features of tonal languages [73]. The focus presents previous efforts on tone recognition for tonal languages such as Chinese, Thai, Vietnamese, Mandarin, Mizo, Bodo, Punjabi and Yoruba. Significantly, the inquiry underscores a notable research gap concerning Manipuri tone recognition. This chapter delves into the progress of tone recognition methodologies across various tonal languages.

3.1 Mandarin Chinese

A technique was developed to recognize lexical tones in Mandarin speech, utilizing vector quantization and Hidden Markov Model (HMM) [49]. The method involved

Table 3.1: An overview of the phonological features of tonal languages

Language Family	Tonal Language	No. of Tones	Types of Tones	Consonants	Vowels
Sino-Tibetan	Mizo	4	falling, high, rising, low	32	5
	Chinese	6	low-falling rising, neutral, high, high-rising, high-falling	19	5
	Mandarin	4	falling-rising, rising, high, falling	22	7
Tibeto-Burman	Bodo	2	high, low	16	6
	Manipuri	2	falling, level	24	8
Tai-Kadai	Thai	5	rising, high, low, mid, falling	20	18
Austro-Asiatic	Vietnamese	6	drop, rising, curve, broken, falling, level	25	11
Indo-European	Punjabi	3	low, mid, high	25	10
Niger-Congo	Yoruba	3	low, mid, high	17	11

extracting features from the fundamental frequency contour and constructing Markov models. Remarkably, the average recognition rates were 91.00% for trisyllabic words, 92.90% for disyllabic words and 97.90% for isolated monosyllabic words. However, the accuracy declined due to tone variation and coarticulation effects. Fu et al. [23] explored using temporal envelope cues in Mandarin Chinese speech recognition. The speech was partitioned into frequency analysis bands and amplitude envelopes were extracted from each band. The findings indicated a gradual increase in recognition rates for vowels, consonants and sentences as the number of bands increased, mirroring trends observed in English speech recognition. Notably, tones were consistently identified at an 80% correct level, significantly influencing sentence recognition. Additionally, a strong correlation between tone recognition and sentence comprehension was observed. In a study by Li et al. [47], the fuzzy C-mean algorithm was utilized for recognizing the four

Mandarin tones. Pitch features were detected using discrete wavelet transform (DWT). The experiment revealed a detection error of 187 syllable characters out of 1177 characters in Mandarin. This approach enhanced accuracy by eliminating false fundamental frequency values and leveraging maximum memberships. Zhang et al. [94] investigated the tone statistics of Mandarin Chinese names and the issue of tone-confusable names in a Chinese voice tag dialing system. Leveraging a Chinese name database comprising 1.6 million names, the investigators devised a tone enhancement method to transform an English speech recognizer into a tonal language speech recognizer. Benchmark testing and pseudo-recognition analysis indicated that augmenting the English version of the speech recognizer with tonal recognition capabilities could enhance the recognition of Chinese names and decrease recognition error rates by 10-20%. Li [48] introduced a speech recognition technique employing Linear Predictive Coding Cepstra (LPCC) vectors for representing each syllable. Utilizing a simplified Bayes decision rule with weighted variance, the method effectively classified 408 speaker-dependent Mandarin syllables, achieving a recognition rate of 94.36% instead of 79.78% obtained using HMM. Notably, the system outperformed other known techniques in speed and accuracy, including the HMM system. Tang and Yin [82] introduced a speaker normalization algorithm to mitigate the impact of Mandarin tone variations on speaker-independent tone recognition systems. They employed a Multi-Layer Perceptron (MLP) model and extracted six features from the Fundamental Frequency (F_0) contour. Tone variations were classified into neutral tone, isolated syllable, disyllable and polysyllable. Notably, the neutral tone, a unique tonal sandhi, typically exhibited low recognition rates in existing systems. When employing these novel methods, experimental findings demonstrated an impressive accuracy rate of 96.64% in Mandarin tone recognition when using these novel methods.

Table 3.2: Work done on tone recognition of Mandarin Chinese

Author and Year	Dataset	Features	Model	Result
Liu et al. 1989 [49]	Collected from 7 male speakers Total no. of utterances - Monosyllabic: 1512 Disyllabic: 1680 Trisyllabic: 1344	F_0 contour	3 HMMs for each type of syllable	Recognition accuracy (%) Monosyllabic: 97.90 Disyllabic: 92.90 Trisyllabic: 91.00
Fu et al. 1998 [23]	Collected from 12 native speakers Test material includes: 21 initial consonants, 35 final vowels, 4 tones, 200 daily-life sentences	Temporal envelop cues	Power-function model	Recognition accuracy: 80%
Li et al. 1999 [47]	1177 Mandarin syllables	F_0 curve fitting	Fuzzy C-means algorithm	Detected error of 187 syllables (including 95 syllables segment error)
Zhang et al. 2002 [94]	1.6 million Chinese names	MFCC	HMM	With information reduce the recognition rate by 10-20%
Li 2003 [48]	408 Mandarin syllables	LPCC	Simplified Bayes decision rule	Recognition accuracy: 94.36%
Tang and Yin 2006 [82]	1254 monosyllabic words, 150 disyllabic words, 123 quadrisyllabic words, 60 passages	6 features derived from F_0	MLP	Recognition accuracy: 96.64%
Xie and Miao 2010 [91]	Collected from 5 male and 6 female speakers. Every speaker pronounced 4 times 1588 syllables with tones	Normalized F_0 curve and energy	MLP	Average recognition accuracy across 4 tones: 93.66%
Wu et al. 2013 [89]	RASC863	Spectral/temporal features, energy contours and pitch contours.	HMM and NN	Recognition accuracy HMM: 60.3% NN: 71.7%

Table 3.2: Work done on tone recognition of Mandarin Chinese (continued)

Author and Year	Dataset	Features	Model	Result
Huang et al. 2017 [37]	863 corpus	LDA MLLT	RNN DNN-RNN GMM-HMM	Recognition accuracy RNN: 70.70% DNN-RNN: 70.40% GMM-HMM: 75.7%
Chao et al. 2019 [11]	863 corpus	Prosodic and articulatory extracted by DBN	SVM	Recognition accuracy: 83.03%
Gao et al. 2019 [25]	SCSC	Mel-spectrogram	ToneNet: CNN Network	Recognition accuracy: 99.16%
Peng et al. 2021 [63]	Hi-Tech Project 863 corpus	Mel-spectrogram cepstrogram	CNN	Recognition accuracy: 88.06%
Yan et al. 2023 [92]	SCSC	FFSs	RF	Recognition accuracy: 93.57%

Xie and Miao [91] introduced an algorithm for speaker normalization and a neural network-based tone recognizer designed specifically for Mandarin. The primary focus was to tackle the challenge posed by the varying tone behavior across different speakers. The system incorporated an enhanced pitch detector and a robust tone identification method to standardize each utterance's F_0 contour. Employing a 15-layer MLP pattern recognizer, the system effectively classified the four tones. The training process utilized a backpropagation rule to minimize the mean squared error between the outputs and targets. Additionally, the paper discussed various tone recognition approaches based on artificial neural networks, including a tone pitch detector and a smoothing scheme. The system achieved an average accuracy rate of 93.66% across all four tones in Mandarin speech recognition.

Wu et al. [89] assessed the efficacy of automatic systems in classifying and recognizing tones from continuous Mandarin Chinese speech. Native Mandarin speakers and two automated methods, neural networks and Hidden Markov Models were employed

in the study. The findings indicate that humans require context to accurately recognize tones, whereas non-native speakers struggle significantly with this task. While machine recognition and human recognition exhibit similar levels of accuracy, they demonstrate distinct error patterns. The study reveals that reasonably accurate tone classification and recognition can be achieved without pitch features. Huang et al. [37] introduced an Encoder-Classifier framework for Mandarin tone modeling employing Recurrent Neural Networks (RNN). This framework extracts features for tone classification and transforms them into fixed-dimensional vectors (tone embeddings) via a softmax layer. The model can be improved through pooling, feature splicing and incorporating syllable-level tone embeddings. Experimental findings demonstrate enhanced tone classification accuracy with the proposed network configurations. The framework exhibits promise for accommodating both sequence-based and segment-based features.

Tone recognition of Chinese syllables was proposed based on the Support Vector Machine (SVM) and Deep Belief Network (DBN) model [11]. The DBN was used to extract the prosodic and articulatory features and SVM was used for the tone classification. The method was applied on the 863-data corpus with an average accuracy rate achieved of 83.03%. A multi-scale model was proposed to capture the features of tone variation of Mandarin affected by complex linguistic and phonetic rules [63]. Cepstrogram and Mel-spectrogram were used as the input features of the speech data and a multi-scale Convolutional Neural Network (CNN) was employed to recognize the tone. Experimental analysis showed the method achieved state-of-the-art results on the Corpus 863 data. Yan et al. [92] proposed a method for Mandarin tone identification based on Random Forest (RF) and three feature fusion sets (FFSs). The Syllable Corpus of Standard Chinese (SCSC), a speaker-independent Mandarin monosyllable corpus, was used to assess the methodology. The findings demonstrate that, even with unbalanced data, the tone recognition algorithm can achieve high tone recognition accuracy and

exhibit strong generalization and classification capabilities. The SCSC dataset is again used to evaluate the model designed using CNN named ToneNet [25], which applies Mel-spectrogram as features, achieving a tone recognition rate of 99.16% accuracy. Table 3.2 shows the summarised work done on tone recognition of Mandarin Chinese.

3.2 Thai

Potisuk et al. [64] proposed a Thai tone recognition system using an extension of Fujisaki's model [24] that incorporates tonal declination and assimilation, achieving an accuracy rate of 89.1%. Maliyaem et al. [51] proposed a method for improving Automatic Speech Recognition (ASR) for continuous Thai words using a combination of MFCC and tone information as input feature vectors. Artificial Neural Network (ANN) multilayer perceptrons are employed to estimate the posterior probabilities of a class model given a sequence of observation inputs. The performance of the ANN method is compared against the Gaussian Mixture Model (GMM) utilized in the Hidden Markov Model Toolkit (HTK). Experimental evaluations involve using 2-gram and 3-gram language models, revealing that the proposed technique effectively enhances Thai-ASR performance by reducing the word error rate.

Theera-Umpon et al. [84] investigated a phoneme recognition system designed for Thai speech, specifically targeting phonemes and tonal accents. The system used discrete Markov models to identify Thai phonemes encompassing initial consonants, vowels and final consonants. A comparison between Mel frequency with perceptual linear prediction and Mel frequency cepstrum coefficients for phoneme recognition purposes is drawn. Neural networks are leveraged to classify tonal accents by analyzing pitch frequency variations across syllables. Encouraging results indicate promising recognition performances for both phonemes and tonal accents. The study suggests

Table 3.3: Work done on tone recognition of Thai language

Author and Year	Dataset	Feature	Model	Result
Potisuk et al. 1999 [64]	Collected from 5 speakers of total 625 utterances	Normalized F_0 contour	Extended Fujisaki Model	Recognition accuracy: 89.10%
Maliyaem et al. 2010 [51]	809 words	MFCC 3 Tone Features	GMM ANN	WER (%) GMM- 2-gram: 26.35 3-gram: 24.76 ANN- 2-gram: 25.20 3-gram: 23.35
Theera-Umpon et al. 2011 [84]	Collected from 20 speakers for speaker-dependent and 10 speakers for speaker-independent of total 2700 syllables	Pitch MFPLP	ANN	Tonal accent accuracy Speaker-dependent: 89.69% Speaker-independent: 89.14%
Srijiranon and Eiamkanitchat 2015 [78]	Collected from 7 male and 7 female speakers, 9 words in different styles	PLP	Neuro-Fuzzy	Accuracy on syllable data Male: 91.25% Female: 84.69% Mix: 78.99%
Kertkeidkachorn et al. 2015 [42]	LOTUS	Energy and Spectral based	HCRF	Recognition Accuracy Energy based: 94.24% Spectral based: 71.01%

that speakers adhering to the pronunciation patterns outlined in the training set tend to achieve superior recognition outcomes. Srijiranon and Eiamkanitchat [78] investigated the performance of the Neuro-fuzzy system as a speech recognition algorithm. Eight Thai words were recorded in different environments, syllables and pronunciations. The aim is to apply this method to instruments for disabled people. Initial findings indicate that each variable exerts distinct influences on recognition accuracy. However, experimental outcomes demonstrate the robustness of the Neuro-fuzzy system against noise, yielding superior recognition rates compared to prevalent algorithms. The research underscores the suitability of the Neuro-fuzzy system for speech recognition applications owing to its heightened classification accuracy. Another study in Thai tone recognition was presented by Kertkeidkachorn et al. [42]. Hidden Conditional Random Field (HCRF) was

used as a tone classifier, applying energy-based and spectral-based features, including Linear Prediction Coefficients (LPC), Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). In isolated words, energy-based systems provide an error rate reduction of 22.40% and in the continuous-speech, PLP-based features achieved an error rate reduction of 13.90%. Table 3.3 shows the summarised work on tone recognition of the Thai language.

3.3 Vietnamese

Nguyen et al. [55] introduced a novel approach to integrating Vietnamese language characteristics into a Large Vocabulary Continuous Speech Recognition System (LVCSR). This method utilizes a HMM for tone recognition and employs a transformation technique to convert monosyllabic words into polysyllabic words. Through adaptation specifically tailored for Vietnamese, experiments conducted on the VNSPEECHCORPUS dataset demonstrated a notable increase in accuracy alongside a remarkable 46% reduction in word error rate. Incorporating polysyllabic words and tone information notably enhances performance by 46%. Nguyen Van Huy [59] introduced a Multi Space Distribution Hidden Markov Model (MSD-HMM) approach tailored for Vietnamese recognition. Their prototype incorporates four independent streams designed to model Vietnamese phonemes alongside tonal information corresponding to syllables. This innovative approach exhibits a notable improvement of 2.49% in accuracy compared to the baseline system. Furthermore, including a tonal phoneme set enhances accuracy by approximately 1% when contrasted with non-tonal phoneme sets. The results underscore the effectiveness of MSD in Vietnamese recognition, showcasing enhancements of 2.49% and 0.54% in accuracy compared to the best baseline system. These findings emphasize the significance of tonal features in recognizing tonal languages. Nguyen et al.

Table 3.4: Work done on tone recognition of Vietnamese language

Author and Year	Dataset	Feature	Model	Result
Nguyen et al. 2008 [55]	VNSpeech corpus	F_0, short-time energy	HMM	46.00% relative reduction of WER
Nguyen Van Huy 2014 [59]	VOV	MFCC, Pitch features using AMDF and NCC	MSD-HMM	Recognition Accuracy: 80.80%
Nguyen et al. 2015 [58]	IOIT2013, VOV, GlobalPhone, VoiceTra, BTEC	BNF	GMM-HMMs	WER (%): 9.14
Nguyen et al. 2016 [56]	VOV	DBNFs + pitch features	LVCSR Neural Network sytem	19.25% improvement observed in tonal phoneme
Nguyen and Cao. 2020 [57]	22 types of voice command of total 2350 recordings	CNN features extraction	SVM CNN	Reduce errors in recognition to- SVM: 7.08% CNN: 5.15%

[58] introduced an algorithm to convert Vietnamese words into tonal phoneme-based pronunciations, resulting in a notable improvement in word error rate (WER) by 1.54%. This approach used a tonal phoneme set to construct acoustic models incorporating tone information and tonal features. Leveraging the Kaldi toolkit, the researchers developed a Vietnamese LVCSR system and extracted bottleneck features from a trained deep neural network tailored for Vietnamese. Incorporating tonal features led to a significant reduction in WER by 4.65%. Nguyen et al. [56] investigated tone modeling in Vietnamese LVCSR systems, focusing on the tonal structure of Vietnamese syllables. Results indicate a notable enhancement in performance, with a 19.25% improvement observed when employing tonal phonemes compared to non-tonal phoneme systems. Additionally, incorporating tonal features as input to neural networks led to an 18% increase in performance. In another study, Nguyen and Cao. [57] introduced a novel method for identifying Vietnamese voice commands by leveraging the results of the

Google Speech Recognition (GSR) service. The study involved the construction of a comprehensive voice command dataset and the development of a correction system employing Support Vector Machine (SVM) and Convolutional Neural Network (CNN) models. The findings underscored the significant enhancement in recognition accuracy achieved by the CNN model, which substantially reduced command error rates from 35.06% to 7.08%. Moreover, the method demonstrated the potential for deploying device control applications via voice command on Android devices. Table 3.4 shows the summarised work done on tone recognition of the Vietnamese language.

3.4 Exploration on other Tonal Languages

Mizo is a tonal language with four tones, is understudied and requires significant resources. To address this, Sarma et al. [68] developed a tone detection method for Mizo using quantitative analysis of acoustic features. Unlike conventional modeling techniques that necessitate extensive datasets for training, their approach introduces a novel method based on slope and height values. Tones are classified into High/Low and Falling/Rising categories, achieving a 70% accuracy rate. This classification relies on pitch values, height and slope parameters, effectively mitigating the consonantal effect by excluding the initial and final 20% of pitch values. Dey et al. [17] created a speech corpus for the under-resourced Mizo language, accompanied by implementing a phone recognition system utilizing advanced acoustic modeling techniques. The corpus comprises passages spoken by 81 native Mizo speakers, encapsulating 36 phonetic units that capture the acoustic nuances of the speech signal. Through exploration, the researchers investigated the efficacy of Hidden Markov Models (HMM) and Deep Neural Networks (DNN) in phone recognition. Results indicate a reduction in phone error rates (PERs) to 13.9% and 15.7% following the integration of language model information.

Incorporating language model data significantly enhances the overall performance of the phone recognition system. Another tone detection of Mizo was proposed using features derived from pitch contour and applied on SVM-based and Deep Neural Network (DNN) based models, achieving an accuracy of 73.39% and 74.11% respectively for the two models [27]. Gogoi et al. [28] proposed a novel approach incorporating Rising tone sandhi alongside four lexical tones, leveraging a 1D-CNN model to learn tone-specific features directly from F0 contours. Experimental findings reveal that the proposed 1D-CNN model achieves an accuracy of 68.18%. Additionally, the study aimed to detect five Mizo tones using a baseline model with six F0 features. Comparative analysis showcases the superiority of the 1D-CNN model over the baseline system, yielding 63% accurate results.

Bhattacharje [4] undertook the recognition of tonal words in the **Bodo** language, aiming to explore modifications to speech recognition systems for incorporating tonal features, often overlooked during system development. Two distinct approaches were proposed: a feature-level solution and a model-level solution. Experimental investigations revealed a significant degradation in speech recognition system performance when applied to tonal words, attributed to removing pitch-related information during feature extraction techniques. Although the addition of prosodic features led to an improvement of nearly 8% in recognition accuracy, overall accuracy remained relatively low. To address this challenge, the utilization of two separate tone recognizers was proposed, resulting in a substantial enhancement in system performance.

ỌDélọBí [95] studied the automatic recognition of Standard **Yoruba** (SY) language tones utilizing Artificial Neural Networks (ANN) and presented a fundamental approach to tone recognition for Yoruba. The models leveraged the fundamental frequency profile of SY syllables to characterize and discriminate three Yoruba tones. Results demonstrated promising performances for Recurrent Neural Network (RNN) and Multi-

layer Perceptron (MLP), with the RNN model achieving higher accuracy rates at 76.00%. A.A. Sosimi and Fakinlede [2] introduced a context-dependent (CD) tone acoustic model tailored for Standard Yorùbá (SY), a language characterized by syllables with tones. The model utilizes the amplitude magnified difference function (AMDF) to derive wide F-contours, automatic syllabification and tri-syllable forced alignment. The training involved the Multi-class Support Vector Machine (MSVM) for tri-tone classification. Results indicated that the MSVM tri-tone system surpassed the hidden Markov model tri-phone system regarding word error rate (WER), achieving an overall accuracy of 97.5678% with optimized parameter tuning through 10-fold cross-validation. Specifically, the MSVM CD tri-tone system exhibited a superior WER of 44.47% compared to the hidden Markov model tri-phone system. However, the study highlights the need for further exploration into the SY CD tone recognition system's ability to model time-sequential nature and handle dialectic variations.

Punjabi is a regional language with variant pronunciations and tones. Conventional research uses basic extraction techniques like Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). Kaur and Singh [41] introduced a novel feature extraction technique called Power-Normalized Cepstral Coefficients (PNCC) for Punjabi ASR. The study employed phone-based modeling to train 158 words in noisy and noise-free environments. Performance analysis revealed an accuracy of 83.05% in noise-free environments and 71.92% in noisy environments. Guglani and Mishra [29] explored enhancements to ASR systems for tonal languages, particularly in Punjabi, emphasizing pitch-dependent features and estimated features of the probability of voicing. Their study evaluated the performance of various pitch features using word error rate (WER). The findings indicated that the Kaldi pitch tracker outperformed other ASR systems, achieving a 1.5% reduction in WER compared to the Yin feature-based ASR. Utilizing the Kaldi pitch tracker facilitated the extraction of tone-related pitch features,

thereby improving the overall performance of the ASR system for Punjabi. Table 3.5 summarised the work done in tonal languages Mizo, Boro, Yoruba and Punjabi.

Table 3.5: Work done on tone recognition of other tonal languages

Author and Year	Language	Dataset	Feature	Model	Result
ỌDélọBí 2008 [95]	Yoruba	Collected from 4 speakers of total 475 data	F_0 curve	MLP RNN	Accuracy MLP: 71.00% RNN: 76.00%
Bhattacharje 2013 [4]	Bodo	Collected from 25 speakers of total 3500 words	MFCC, Prosodic	Single RNN based, Separate RNN based	Accuracy Single RNN: 74.29% Separate RNN 83.57%
Sarma et al. 2015 [68]	Mizo	Collected from 5 Mizo speakers of 672 high tone, 2254 low tone, 808 falling and 650 rising tones	pitch contour	Classify based on variance, height and slope of pitch	Accuracy: 70.00%
Kaur and Singh 2016 [41]	Punjabi	150 distinct words Noise-free: 16 speakers of 700 audio files Noisy: 12 speakers 672 files	PNCC	HMM	Accuracy (%) Noise-free: 83.05 Noisy: 71.92
Dey et al. 2017 [17]	Mizo	Collected from 81 native speakers, 9 passages, 38 sentences/ passage	MFCC	DNN SGMM	PER reduces to DNN: 13.9% SGMM: 15.7%
A.A. Sosimi and Fakinlede 2019 [2]	Yoruba	RBCOM SY database	F_0	MSVM	Accuracy: 97.56%
Gogoi et al. 2020 [27]	Mizo	Collected from 19 native speakers of 31950 syllables	F_0 contour	SVM DNN	Recognition Accuracy SVM: 73.39% DNN: 74.11%

Table 3.5: Work done on tone recognition of other tonal languages (continued)

Author and Year	Language	Dataset	Feature	Model	Result
Guglani and Mishra 2020 [29]	Punjabi	Collected from 100 speakers, total of 12.5 h	Kaldi pitch, F_0 variation	Kaldi Decoder	WER reduces by 1.5%
Gogoi et al. 2021 [28]	Mizo	Collected from 19 speakers of 1.8 h	F0 contour	1D-CNN	Accuracy: 68.18%

3.5 Research on Manipuri in Various Domains

This section provides an overview of the research and contributions to studying the Manipuri language across various domains. A comprehensive review highlights the diverse scholarly endeavors and initiatives that explore and promote the Manipuri language and its associated heritage.

Kh Raju Singha and Singha [43] introduced a Manipuri Part-of-Speech (POS) tagging model utilizing Hidden Markov Models (HMM). They generated a small set of tagged sentences from the Manipuri Rule-based Tagger. This system has potential applications in Manipuri language processing, speech recognition and machine translation. Furthermore, 80% of the sequences generated automatically for the test set were correct compared to the manually tagged result of those sentences. Sagolsem Poireiton Meitei [66] developed an Electronic Manipuri Bilingual dictionary, combining English-Manipuri lexicographic data and conceptual information. The user-friendly, vibrant and robust JAVA-MYSQL database-based dictionary streamlines teaching and learning, capturing students' interest and aiding instructors in achieving pedagogical objectives.

Table 3.6: Work done on Manipuri in different domain

1 **Author and Year**	**Domain**	**Method**	**Result**
Kh Raju Singha and Singha 2012 [43]	Speech Tagging	HMM	Accuracy: 80.00%
Sagolsem Poireiton Meitei 2012 [66]	Electronic Bilingual Dictionary	Implement using JAVA and MYSQL	Developed Electronic-Dictionary
Kishorjit Nongmeikapam 2014 [45]	Speech Tagging	POS tagging using CRF	Recall: 71.30% Precision: 77.36% F-measure: 74.21%
Gyanendro Singh et al. 2016 [32]	Automatic Syllabification	EPS	Accuracy: 98.00%
Dutta et al. 2017 [19]	Phonetic Engine	HMM using MFCC, LPCC, PLP	PLP and MFCC are superior to LPCC
Patel et al. 2018 [62]	ASR and KWS	DNN-HMM	WER: 13.57% EER: 7.64%
Gha 2021 [26]	Language Identification	PPRLM, HMM	Accuracy: 95.00%
Sanayai Meetei et al. 2021 [67]	Speech to Text Translation System	GMM-HMM, TDNN	WER: 2.53%
Gurumayum Robert Michael 2022 [31]	Emotion Recognition	CNN	Accuracy: 71.00%

Kishorjit Nongmeikapam [45] explored the chunking process of Manipuri, a notably agglutinate language. Employing Conditional Random Field (CRF) for Part of Speech (POS) tagging, their system achieved a refined Manipuri text with a recall rate of 71.30%, precision of 77.36% and an F-measure of 74.21%. Gyanendro Singh et al. [32] presented data-driven methods for automatically syllabifying words in the Manipuri language using Entropy Based Phonotactic Segmentation (EPS). The methods include a language-independent rule-based approach, sequence labeling and a hybrid approach. Experimental observations show that the proposed methods outperform the baseline rule-based method, with a word accuracy of 98%.

Dutta et al. [19] evaluated a Phonetic Engine (PE) designed for the Manipuri language, employing phonetic transcriptions and Hidden Markov Model (HMM) modeling. The study utilized symbols from the International Phonetic Alphabet (IPA) and a 5-

state HMM featuring 32 mixtures in each state. They analyzed the performance of the Manipuri PE across three widely used spectral features: Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) and Linear Predictive Coding Cepstral Coefficients (LPCC), with dimensions of 13, 26 and 39 coefficients. The results indicated that PLP and MFCC outperformed LPCC across all conditions.

Patel et al. [62] engineered an ASR and Keyword Search (KWS) system tailored for Manipuri. Leveraging over 90 hours of transcribed speech data contributed by over 300 speakers, the team employed the KALDI speech recognition toolkit for system development. The Manipuri ASR system and KWS have been seamlessly integrated into a visual interface for demonstrative purposes. Using a DNN-HMM model, the system achieved remarkable performance metrics, boasting a Word Error Rate (WER) of 13.57% for ASR and an Equal Error Rate (EER) of 7.64% for KWS. Another work by Patel et al. [61] introduced an extensive methodology for building speech technology applications tailored to Manipuri, encompassing Language Identification, Speech-to-Text, Keyword Search and Speaker Diarization frameworks. The platform accommodates Manipuri and English languages with the potential for extension to other linguistic domains. A user-friendly interface facilitates demonstration, while future developments aim to integrate diarization with Speaker Identification and Machine Translation to support native English language functionality. Gha [26] introduced a spoken language identification system for under-resourced languages such as Hindi and Manipuri. The system demonstrates accurate language identification capabilities and undergoes testing for real-time identification scenarios. Phonetic tokenization and acoustic modeling are employed in the system's front-end and back-end processes. Furthermore, integrating acoustic and language models enhances the system's performance for real-time identification tasks.

Sanayai Meetei et al. [67] developed speech-to-text translation systems for Manipuri-English. A dataset comprising a Manipuri-English parallel corpus alongside its audio

version is established. Benchmark assessments are conducted using two methodologies: pipeline models integrating ASR and Machine Translation (MT) and end-to-end translation employing the Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) and Time Delay Neural Network (TDNN) Acoustic models. Both approaches surpass the end-to-end translation model. The experimental findings reveal that the TDNN model exhibits superior performance over the GMM-HMM model by a considerable margin of 2.53% Word Error Rate (WER).

Emotion recognition is vital in enhancing Human-Computer Interaction systems and finding applications across diverse domains such as education, gaming, medicine and automobiles. Gurumayum Robert Michael [31] conducted a study where they trained a Convolutional Neural Network (CNN) model using a Manipuri speech dataset to identify four primary emotions. Their model achieved an initial accuracy of 46%, which increased to 71% with data augmentation techniques. Table 3.6 summarised the work done on Manipuri in different domains.

3.6 Research on Manipuri Tones

Research on Manipuri language and phonology has yielded valuable insights into its **tonal system** and linguistic structure. Thoudam's doctoral thesis [85] includes a chapter dedicated to Manipuri phonology, proposing the existence of only two distinctive tones: falling and level tones. L. [46] similarly argued for the presence of two tones, falling and level, in his master's thesis. Chelliah [12] studied the morphology and phonology of Manipuri, explaining several phonological rules. In 1997, Chelliah [13] clarified the tone system in Manipuri, introducing a framework that accurately described Manipuri's two-way tonal contrast: low tone and high tone. In her experiments, she utilized fundamental frequency contours as phonetic representations of the underlying tone pattern. Devi [16]

examined loanwords (foreign words) in Manipuri and their impact, providing insights into Manipuri phonology, which consists of six vowels, six diphthongs, twenty-four consonants and two tones. She observed that Manipuri loanwords primarily originate from Indo-European languages such as Sanskrit, Bengali and Hindi, which have altered the phonological structure of Manipuri. Singh et al. [76] addressed issues related to Meiteiron writing using the Bengali script and advocated using the Meitei Mayek Script to resolve orthographic challenges. Singh and Singh [75] investigated the phonemes in Manipuri, identifying 24 consonant phonemes, with 15 indigenous consonant phonemes occurring in various syllable positions (initial, central and final) and 9 exotic phonemes used exclusively in central positions. Shastri and Uppunda [70] delineated the attributes of lexical tones in monosyllabic and bisyllabic words in Manipuri through acoustic analysis. They utilized pitch contours to discern the tones and observed variations in pitch contours for the same words across different speakers. Nevertheless, despite these variations, the meaning of the words remained correctly identified. The researchers posited that listeners engaged in the normalization of lexical tones produced by diverse speakers.

However, despite extensive linguistic research, work has yet to be done on the tone recognition of Manipuri, indicating a gap in exploring its tonal characteristics and highlighting an area ripe for further investigation.

4

Development of ManiTo: A Manipuri Tonal Contrast Dataset

Tone is a defining feature of tonal languages, encompassing the pitch patterns that carry meaning distinctions within words. In tonal languages, such as Manipuri, Mizo, Punjabi, Mandarin Chinese, Thai, Vietnamese and Yoruba, variations in tone can change the meaning of a word, leading to a system where the same sequence of vowels and consonants can have different meanings based solely on the pitch contour. Typically, tonal languages utilize distinct pitch patterns applied to syllables or words to differentiate lexical or grammatical meanings. These tones can vary in number and complexity across different tonal languages. The study of tones in tonal languages involves analyzing pitch contours, understanding tone types and investigating how tones interact with other linguistic features. These languages pose unique challenges and opportunities for linguistic research and technology development. One crucial aspect of advancing our understanding and leveraging the potential of tonal languages is the development of comprehensive datasets tailored to their specific linguistic characteristics. Creating comprehensive tonal language datasets can contribute significantly to linguistic research, language preservation and technological innovation in speech and language processing.

This chapter details the development of a tonal contrast word pair dataset for Manipuri, ManiTo. It is imperative to have a well-constructed dataset demonstrating tonal contrasts in words to study and develop a tone recognition system for Manipuri. Af-

ter extensive search and investigation, it has been discovered that no such dataset is available for Manipuri. Therefore, we have taken the initiative to develop the dataset.

4.1 Description of Data

Manipuri is a tonal language predominantly spoken in Manipur, a northeastern state of India and in a few neighboring countries. Its script is known as Meitei Mayek. The language consists of two tones: the level and the falling tones. Each syllable in Manipuri is assigned one of these two tones. In written transcripts, the level tone is left unmarked, whereas the falling tone is indicated by a diacritical mark lum mayek (·), placed after the syllable carrying the falling tone.

Understanding tones of tonal languages requires in-depth studies of minimal pairs of words that differ just by pitch. Pitch is a perceptual quality of sound that correlates with the vibration of the vocal cord. It carries the meaning of the word in tonal languages. Fifty-four tonal contrast word pairs are collected from different sources [74, 18, 13, 80, 85, 44] to investigate the two-way contrastive tones. Each word pair is meticulously selected to highlight specific tonal patterns and contrasts, providing valuable insights into the tonal system of the language. The words and their corresponding meanings are listed in Table 4.1.

Table 4.1: List of Manipuri tonal contrast word pairs with their respective meanings

Sl. No.	Falling Tone	Meaning	Level Tone	Meaning
1	ꯎꯟ·ꯆ /ùn/	skin	ꯎꯟꯆ /un/	ice, snow
2	ꯏꯅ· /ìn/	push	ꯏꯅ /in/	wrap
3	ꯇꯥꯡ· /tʰòŋ/	door	ꯇꯥꯡ /tʰoŋ/	bridge
4	ꯃꯤ· /mì/	person	ꯃꯤ /mi/	spider

Table 4.1: List of Manipuri tonal contrast word Pairs with their respective meanings (continued)

Sl. No.	Falling Tone	Meaning	Level Tone	Meaning
5	ꯄꯨ꯭ꯕ꯭ /pùbə/	to borrow	ꯄꯨꯕ꯭ /pubə/	to bring
6	ꯀ꯭ /cà/	wax	ꯀ꯭ /ca/	tea
7	ꯏ꯭ /ì/	blood	ꯏ꯭ /i/	thatch
8	ꯀ꯭ꯍꯏ꯭ /kʰòi/	navel	ꯀ꯭ꯍꯏ꯭ /kʰoi/	bee
9	ꯂ꯭ /là/	wide basket	ꯂ꯭ /la/	banana leaf
10	ꯁꯤꯡ꯭ /sìŋ/	firewood	ꯁꯤꯡ꯭ /siŋ/	ginger
11	ꯁꯃ꯭ /sə̀m/	hair	ꯁꯃ꯭ /səm/	basket
12	ꯃꯡ꯭ /mə̀ŋ/	dream	ꯃꯡ꯭ /məŋ/	grave
13	ꯀꯥꯡ꯭ /kàŋ/	mosquito	ꯀꯥꯡ꯭ /kaŋ/	Manipur game
14	ꯇꯤꯟ꯭ /tìn/	worm	ꯇꯤꯟ꯭ /tin/	saliva
15	ꯄꯤ꯭ /pì/	give	ꯄꯤ꯭ /pi/	tear
16	ꯀꯍꯡ꯭ /kʰòŋ/	leg	ꯀꯍꯡ꯭ /kʰoŋ/	canal
17	ꯀꯥꯟꯕ꯭ /kànbə/	hard	ꯀꯥꯟꯕ꯭ /kənbə/	protect
18	ꯇꯥ꯭ /tàbə/	to hear	ꯇꯥ꯭ /tabə/	to fall
19	ꯀꯥ꯭ /càbə/	eating	ꯀꯥ꯭ /cabə/	swim visibly
20	ꯀꯝ꯭ /kə̀mbə/	killing	ꯀꯝ꯭ /kəmbə/	over ripen
21	ꯂꯩꯡ꯭ /lə̀ŋbə/	movement	ꯂꯩꯡ꯭ /ləŋbə/	making row
22	ꯑꯁꯡ꯭ /asə̀ŋbə/	to dye	ꯑꯁꯡ꯭ /asəŋbə/	green color
23	ꯀꯥ꯭ /kàbə/	to climb	ꯀꯥ꯭ /kabə/	burnt

Table 4.1: List of Manipuri tonal contrast word Pairs with their respective meanings (continued)

Sl. No.	Falling Tone	Meaning	Level Tone	Meaning
24	ꯇꯣꯡ꯭ꯕ /tòŋbə/	raised platform	ꯇꯣꯡꯕ /toŋbə/	riding
25	ꯇꯨꯝ꯭ꯕ /tùmbə/	sleep	ꯇꯨꯝꯕ /tumbə/	diluted
26	ꯂꯨ꯭ꯕ /lùbə/	clear	ꯂꯨꯕ /lubə/	difficult
27	ꯁꯥ꯭ꯕ /sàbə/	hot	ꯁꯥꯕ /sabə/	to make
28	ꯂꯥꯎ꯭ꯕ /laùbə/	wide open	ꯂꯥꯎꯕ /laubə/	shout
29	ꯀꯥꯎ꯭ꯔ꯭ꯕ /kàwbə/	to kick	ꯀꯥꯎꯔꯕ /kawbə/	to forget
30	ꯍꯥꯏ꯭ꯕ /haìbə/	to say	ꯍꯥꯏꯕ /haibə/	to swing
31	ꯀꯪ꯭ꯕ /kaùbə/	short	ꯀꯪꯕ /kaubə/	call
32	ꯂꯣꯡ꯭ /lòŋ/	spear	ꯂꯣꯡ /loŋ/	finishing net
33	ꯄꯥ꯭ꯕ /pàbə/	thin	ꯄꯥꯕ /pabə/	read
34	ꯇꯨ꯭ꯕ /tùbə/	tailoring	ꯇꯨꯕ /tubə/	fall
35	ꯁꯤ꯭ꯕ /sìbə/	break down	ꯁꯤꯕ /sibə/	death
36	ꯁꯣꯏ꯭ꯕ /soìbə/	mistake	ꯁꯣꯏꯕ /soibə/	cutting
37	ꯂꯦ꯭ꯕ /làbə/	vivid	ꯂꯦꯕ /labə/	male
38	ꯂꯦꯟ꯭ꯕ /lànbə/	crossing	ꯂꯦꯟꯕ /lanbə/	wrong
39	ꯀꯣꯏ꯭ꯕ /koìbə/	travelling	ꯀꯣꯏꯕ /koibə/	long way
40	ꯂꯨꯝ꯭ꯕ /lùmbə/	heavy	ꯂꯨꯝꯕ /lumbə/	warm
41	ꯁꯣꯟ꯭ꯕ /sònbə/	weak	ꯁꯣꯟꯕ /sonbə/	chanting
42	ꯆꯑꯁꯤ꯭ /cʰàniŋ/	lower abdomen	ꯆꯑꯁꯤ /cʰəniŋ/	measuring unit

Table 4.1: List of Manipuri tonal contrast word Pairs with their respective meanings (continued)

Sl. No.	Falling Tone	Meaning	Level Tone	Meaning
43	ꯁꯨ·ꯕ /sùbə/	washing	ꯁꯨꯕ /subə/	hammering
44	ꯃꯨ·ꯕ /mùbə/	roast	ꯃꯨꯕ /mubə/	black
45	ꯄꯩꯕ·ꯕ /paìbə/	touch	ꯄꯩꯕꯕ /paibə/	fly
46	ꯄꯨꯃ· /pùŋ/	drum	ꯄꯨꯃ /puŋ/	hour
47	ꯀꯪ· /kài/	granny	ꯀꯪ /kəi/	tiger
48	ꯍꯨꯕ· /hùi/	dog	ꯍꯨꯕ /hui/	rivet
49	ꯂꯪ· /lài/	purchase	ꯂꯪ /ləi/	tongue
50	ꯄꯦ· /pè/	grimace	ꯄꯦ /pe/	umbrella
51	ꯂꯤ· /lì/	to narrate	ꯂꯤ /li/	old
52	ꯊ· /lù/	fishing trap	ꯊ /lu/	head
53	ꯋꯥ· /wà/	word	ꯋꯥ /bamboo/	old
54	ꯃ· /mà/	bedbug	ꯃ /ma/	third person

4.2 Speakers' Information

The dataset comprises recordings from 20 native speakers, consisting of 11 males and 9 females, aged between 19 to 70 years. All speakers are native Manipuri speakers, with three of them (two males and one female) employed in the Linguistic Department of Manipur University, Imphal. Their recordings were conducted in the Audio, Visual, Language and Phonetic Laboratory Complex of Manipur University. The remaining

recordings were performed in a small studio, a quiet office and a home environment. Additionally, two speakers are employed at a radio station, known for their excellent pronunciation skills, while others hold professorships in the Manipuri language. The details of the speakers and their recording locations are provided in Table 4.2.

Table 4.2: Details of the speakers and the locations where the recordings were conducted

Speaker ID	Gender	Age	Recording Place	Occupation
1	Male	34	Audio, Visual, Language and Phonetic Laboratory Complex of Manipur University	Ph.D. Scholar, Linguistic Dept., Specialist in Linguistics
2	Female	32	Audio, Visual, Language and Phonetic Laboratory Complex of Manipur University	Ph.D. Scholar, Linguistic Dept., Specialist in Linguistics
3	Male	45	Audio, Visual, Language and Phonetic Laboratory Complex of Manipur University	Junior Research Officer, Doctorate
4	Female	29	Quiet room in the speaker's residence at Imphal	Lecturer in BCA Dept.
5	Male	19	Quiet room in the speaker's residence at Imphal	College Student
6	Female	27	Quiet room in the speaker's residence at Imphal	Ph.D. Scholar in Speech Recognition
7	Male	29	Quiet office room at the IIT Guwahati	Ph.D. Scholar in Center of Linguistics.
8	Male	29	Quiet office room at the IIT Guwahati	Ph.D. Scholar in EEE Dept.
9	Male	29	Quiet office room at the IIT Guwahati	Ph.D. Scholar in EEE Dept.
10	Female	55	Quiet Home Studio at Imphal	Experienced in Hosting Radio Shows, good pronunciation skills
11	Female	56	Quiet Home Studio at Imphal	Transcribing Officer
12	Male	55	Quiet Home Studio at Imphal	Experienced in Hosting Radio Shows, good pronunciation skills
13	Male	55	Quiet Home Studio at Imphal	Assistant Professor
14	Female	63	Quiet Home Studio at Imphal	Retired Lecturer
15	Male	55	Quiet Home Studio at Imphal	Businessman
16	Male	56	Quiet Home Studio at Imphal	Businessman
17	Female	49	Quiet Home Studio at Imphal	Assistant Professor

Table 4.2: Details of the speakers and the locations where the recordings were conducted (continued)

Speaker ID	Gender	Age	Recording Place	Occupation
18	Female	26	Quiet room in the speaker's residence at Imphal	Master's Student
19	Female	64	Quiet room in the speaker's residence at Imphal	Retired Teacher
20	Male	70	Quiet room in the speaker's residence at Imphal	Retired Librarian

4.3 Recording Software and Set-up

A total of 54 tonal contrast words, featuring five to ten instances of each pair with pauses between word utterances, were recorded individually for each participant. A laptop microphone (MMDEVAPI/AudioEndpoints) and a Logitech headset (H151) were used to record the participants' speech. The Cool Edit 2000 (v1) [79] tool records the utterances. While recording, the following three parameters have been set in Cool Edit 2000:

- *Sampling rate:* The number of samples per second to be captured by the microphone into the system. The sampling rate is set to 44,100 Hz.
- *Channel:* Mono channel is selected. In mono, all audio signals are routed through a single audio channel.
- *Resolution:* Each sample is represented using 16 bits.
- *Format of saving the audio:* .wav

4.4 Annotation

Data annotation is essential for converting raw data into actionable insights and advancing sophisticated machine learning algorithms. A meticulous manual analysis

Fig. 4.1: Recording scenes in the laboratory and quiet home studio environment

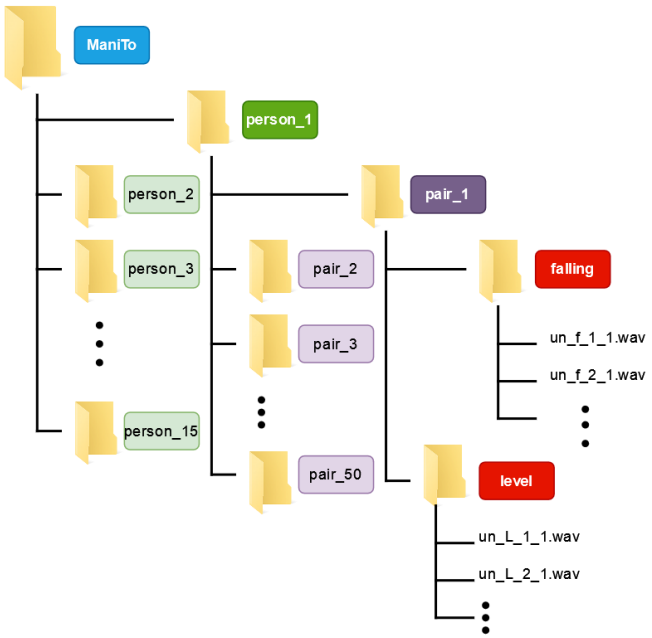


Fig. 4.2: ManiTo directory structure

and segmentation process is employed in the context of the recorded speech sounds. Approximately, 1000 silence samples are added at the beginning and end of each word to ensure comprehensive coverage. Each utterance is then meticulously labeled, including details such as:

- Word name
- Tone type: L for level tone and f for Falling tone
- Utterance number: 1 to 10
- Speaker ID: 1 to 20

For instance, the file 'un_f_2_1.wav' corresponds to the annotation, word: un, tone: falling, utterance no.: 2, speaker ID: 1.

The labeled isolated words are subsequently stored in the designated directory, following a '.wav' format, as illustrated in Fig. 4.2. This structured annotation process

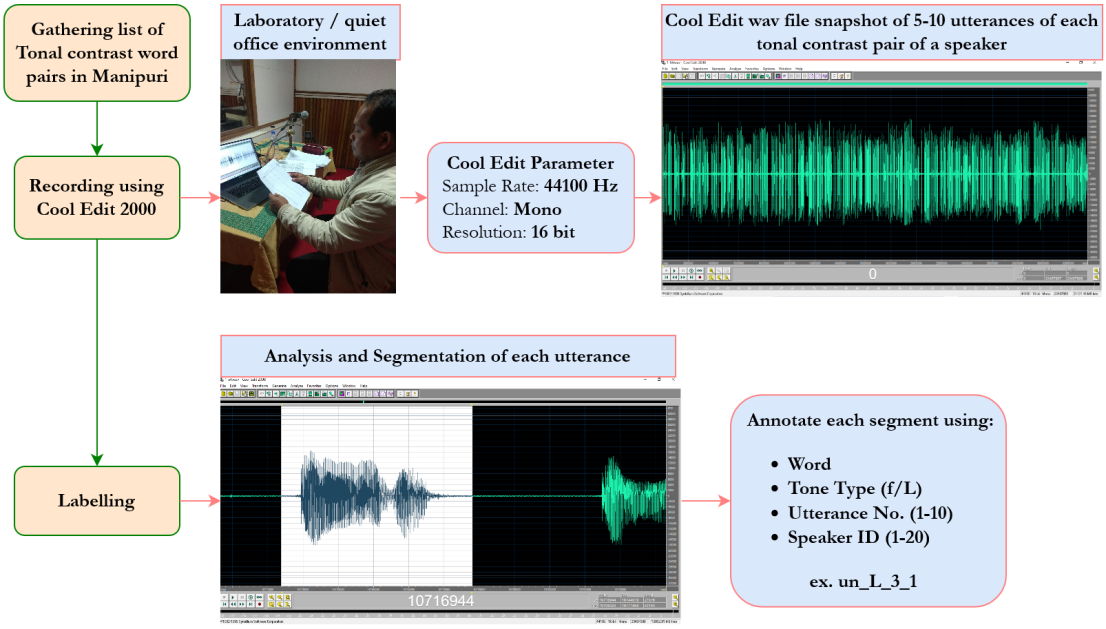


Fig. 4.3: Flowchart showing the creation of ManiTo

enhances the data's richness and relevance, contributing to the effectiveness of subsequent machine-learning endeavors. The flowchart for building the dataset ManiTo is shown in Fig. 4.3.

4.5 Variants of ManiTo

Working directly with raw audio data in deep learning poses challenges due to its complex and high-dimensional nature. To address this, a common practice is transforming the raw audio signals into a 2-dimensional representation, such as spectrograms. These representations capture the frequency and temporal characteristics of the audio, making it more amenable to processing by standard convolutional neural network (CNN) architectures, which are typically designed for image data. We have developed two variants of the audio dataset ManiTo to accommodate this approach. The following subsections explore these two variants of ManiTo in detail.

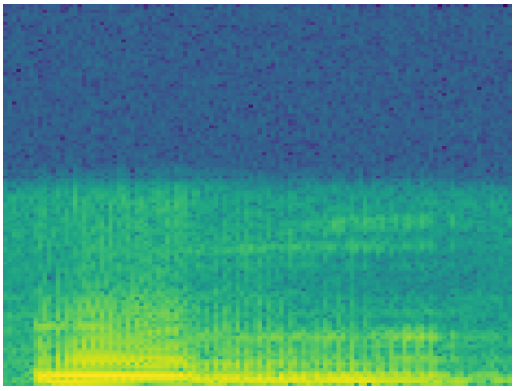


Fig. 4.4: Spectrogram of Falling Tone 'un'

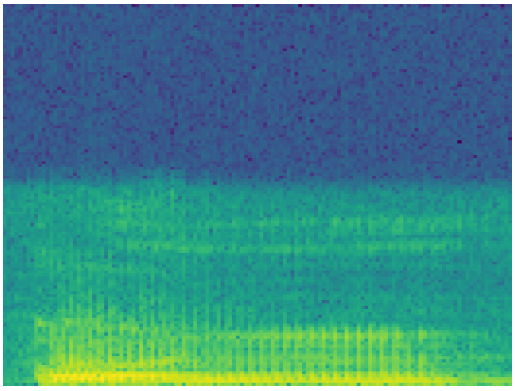


Fig. 4.5: Spectrogram of Level Tone 'un'

4.5.1 ManiToSpc: Spectrogram of ManiTo

A spectrogram visually represents a signal's strength over time at various frequencies present in a waveform. Widely used in audio signal processing, voice analysis and music analysis, spectrograms offer insight into how the frequency components of a sound evolve over time. Given the complexity of raw audio data, many deep learning techniques opt to convert it into a 2D representation, often using spectrograms. This conversion allows for the utilization of standard convolutional neural network (CNN) architectures, which are adept at processing image data. Spectrograms offer advantages over traditional handcrafted features by preserving more audio signal information while maintaining a lower-dimensional format [90]. The audio files of ManiTo are converted into their respective spectrogram [38], saved in PNG format and labeled the file with the word name, tone type, utterance number, speaker ID and sp denoting spectrogram, for example, un_f_1_12_sp.png where this png file contains the spectrogram of the utterance 'un,' falling tone, utterance 1 spoken by speaker ID 12. Fig. 4.4 and Fig. 4.5 show spectrograms of the falling and level tone words 'un' spoken by speaker ID 1. Finally, ManiToSpc consists of labeled spectrogram images of tonal contrast words of Manipuri for 2.18 GB.

4.5.2 ManiToScl: Scalogram of ManiTo

Scalogram is another representation of visualizing signals. It is the absolute coefficient value of a signal's Continuous Wavelet Transform (CWT). Transforming a one-dimensional signal from the time domain to the frequency domain using a wavelet helps analyze the high-resolution signal. By visually representing signals at various scales and frequencies through CWT, hidden features can be detected in the frequency-time domain. These features can adopted in machine learning and deep learning models [33]. Scalogram of ManiTo wav files are obtained using the python package ssqueezepy (v 0.6.3) [53], saved in png format and labeled with word name, tone type, utterance number, speaker ID and sc denoting scalogram, for example, un_f_3_5_sc.png, where this png file contains the scalogram image of word un, falling tone, third utterance, spoken by speaker ID 5. Fig. 4.6 shows the scalogram of the falling tone word 'un' and Fig. 4.7 shows the scalogram of the level tone word 'un' spoken by speaker ID 1. Finally, ManiToScl consists of labeled scalogram images of tonal contrast words of Manipuri, a total of 2.05 GB of data.

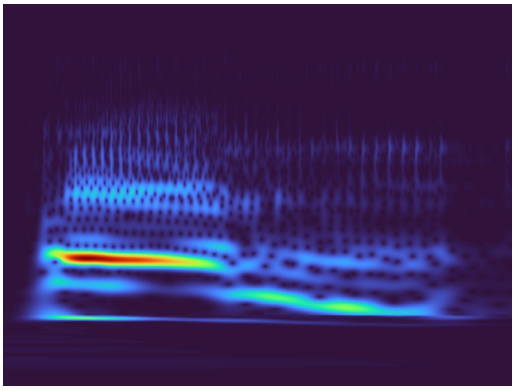


Fig. 4.6: Scalogram of Falling Tone 'un'

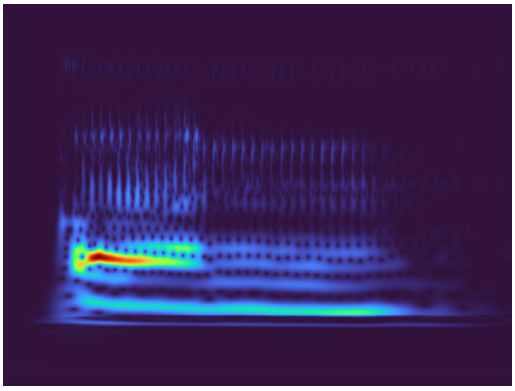


Fig. 4.7: Scalogram of Level Tone 'un'

Table 4.3: The count of utterances made by each speaker and their ages

Speaker ID	Gender	Age	No. of utterances	Speaker ID	Gender	Age	No. of utterances
1	Male	34	524	11	Female	56	1033
2	Female	32	474	12	Male	55	1041
3	Male	45	515	13	Male	55	1029
4	Female	29	513	14	Female	63	989
5	Male	19	538	15	Male	55	970
6	Female	27	533	16	Male	56	1010
7	Male	29	1241	17	Female	49	727
8	Male	29	1188	18	Female	26	1119
9	Male	29	1234	19	Female	64	1080
10	Female	55	1059	20	Male	70	1020

4.6 Conclusion

The developed speech corpus ManiTo is a handcrafted database comprising 17,837 samples of isolated Manipuri tonal contrast words from 20 native speakers of size 1.03 GB. Table 4.3 presents the number of utterances made by each speaker and their corresponding ages. Two variants of the dataset have been created. The first variant, ManiToSpc, comprises spectrograms extracted from the WAV files in ManiTo, providing a two-dimensional representation of the audio data. The second variant, ManiToScl, consists of scalograms generated from the audio files in the ManiTo corpus. These representations pave the way for adopting deep learning techniques that treat audio data as images, offering future research and analysis opportunities in audio processing and classification.

Developing the dataset for Manipuri tones marks a significant step towards facilitating research and technology initiatives in understanding and processing the Manipuri language. Through meticulous recording, annotation and organization, the dataset now offers a valuable resource for studying tonal contrasts and patterns within Manipuri speech. Including speakers of diverse ages and recording environments enhances the

dataset's robustness and applicability across various linguistic contexts. Researchers and developers can leverage this dataset to advance speech recognition, language processing and linguistic studies focused on Manipuri tones. Moving forward, continuous updates and expansions to the dataset will ensure its relevance and effectiveness in supporting future research and technological advancements in Manipuri language processing.

5

Analysis of Manipuri Tones in ManiTo: A Tonal Contrast Database

The analysis of tone in Manipuri is crucial for several reasons. Firstly, Manipuri is a tonal language, meaning the tone of words can change meanings. Understanding the tones used in Manipuri is essential for accurate communication and comprehension among speakers. Secondly, tone analysis in Manipuri aids in linguistic research and documentation. It allows linguists and language scholars to study the phonological characteristics of the language, including its tonal system and patterns. This research contributes to the preservation and documentation of Manipuri culture and heritage. Moreover, tone analysis is valuable for language education and literacy programs in Manipuri-speaking communities. By understanding and teaching the tones of the language, educators can improve language learning outcomes and promote linguistic diversity. Additionally, tone analysis in Manipuri is relevant for developing language technologies and tools. Speech recognition systems, machine translation algorithms and other language processing applications can benefit from accurate tone analysis to improve their performance in understanding and generating Manipuri speech. Overall, the importance of tone analysis in Manipuri lies in its role in communication, linguistic research, education and technological development within Manipuri-speaking communities and beyond.

5.1 Work on Manipuri Tones

Research on the Manipuri language and its phonology has provided significant insights into its tonal system and linguistic structure. Thoudam's doctoral thesis [85] dedicated a chapter to Manipuri phonology, proposing the existence of only two distinctive tones: falling and level tones. Similarly, Mahabir [46] argued for the presence of two tones, falling and level, in his master's thesis. In 1997, Chelliah [13] clarified the tone system in Manipuri, introducing a framework that accurately described Manipuri's two-way tonal contrast: low tone and high tone. In her experiment, she utilized fundamental frequency contours as phonetic representations of the underlying tone pattern. Meiraba [80] asserted that the rhyme in Manipuri serves as the tone-bearing unit. The paper reported Manipuri's relatively simple tonal system results from a rich consonantal inventory that can occur in the coda position and impact how tonal contrast is perceived. Yashawanta [74] categorized the two tones of Manipuri as level and falling, presuming that the tonal feature was gradually losing prominence. Automated Speech Recognition (ASR) plays a pivotal role in supporting language learning by providing feedback on pronunciation and intonation. Incorporating tonal information is crucial for tonal languages like Manipuri to build robust ASR systems. Consequently, identifying features that can effectively disambiguate between the two tones of Manipuri becomes imperative. Furthermore, ASR aids in documenting endangered languages by recording and transcribing spoken language, creating a valuable resource for linguists, anthropologists and other researchers interested in language and culture.

5.2 ManiTo: A Tonal Contrast Database

ManiTo is a comprehensive isolated tonal contrast speech database explicitly designed to investigate the Manipuri language's two tones, level and falling. It comprises 54

tonal contrast word pairs collected from 20 native speakers, including 11 males and 9 females aged 19 to 70. The samples are recorded with 16-bit resolution, a sampling rate of 44,100 Hz, using a mono channel utilizing Cool Edit 2000 (v1) [79]. Each recorded speech segment is meticulously analyzed, manually segmented and labeled with the corresponding word name, tone type, utterance number and speaker ID. Presently, the dataset encompasses 17,837 samples of labeled speech data, providing a valuable resource for further research and analysis in studying Manipuri tonal contrasts and related linguistic phenomena.

5.3 Amplitude Normalization

Amplitude normalization is a critical process in speech analysis to ensure consistent and balanced audio loudness across various speakers and recordings. As outlined in Algorithm 1, this normalization technique is integral to extracting relevant features from speech samples and minimizing the audio signal's dynamic range. In the context of ManiTo's speech samples, which were recorded at a sampling rate of 44.1kHz using 16-bit Linear Pulse Code Modulation (LPCM), the signal's magnitude typically spans from -32768 to +32767, with the peak magnitude set at a level of 10,000. The goal of applying normalization to all speech samples is to standardize the amplitude levels across different recordings and speakers. The process involves adjusting the amplitudes to a uniform scale, facilitating consistent analysis and interpretation of the speech data.

Algorithm 1 describes the process of amplitude normalization for an audio sample. Initially, the audio file is read (line 2) and its data is extracted using the Linear Pulse Code Modulation (LPCM) technique, accomplished with the scipy.io.wavfile.read function available in the SciPy [40] library for Python. This function enables the retrieval of sample values from the WAV file. The algorithm then identifies the highest value within

Algorithm 1: Amplitude Normalization of Audio Sample

Input: $audio$: audio file
Output: $norm\_wav$: amplitude normalized audio sample

```
1 begin
2     in_wav ← ReadAudio(audio)
3     smp_value ← GetDataLPCM(in_wav)
4     peak_value ← GetHighest(smp_value)
5     norm_value ← 10000/ peak_value
6     for value in smp_value do
7         norm_data ← value * norm_value
8     norm_wav ← ConvertToAudio(norm_data)
9     return norm_wav
```

the sample (line 4), representing the peak magnitude of the audio signal. Subsequently, a normalization factor is computed by dividing 10,000 by the peak value. This factor is used to scale each sample value in the audio file to ensure that the peak magnitude reaches the desired level of 10,000. The algorithm iterates through each sample value, multiplying it by the normalization factor to obtain the normalized data. Finally, the normalized data is converted into an audio file format and the resulting normalized audio sample, norm_wav, is returned.

Figure 5.1 visually represents the waveform of speech samples captured from two speakers before and after the normalization process. This illustration highlights the successful balancing and equalization achieved through amplitude normalization. Speaker 1 is noted for having a loud voice, with recordings conducted in a laboratory environment, while Speaker 7 possesses a softer voice, with recordings captured in a quiet office setting. Following the amplitude normalization procedure, both speakers' recordings are adjusted to reach the same level, effectively mitigating the volume disparity between the speakers and environments.

This algorithm effectively standardizes the amplitude levels of the audio sample, ensuring consistent loudness levels across different recordings and speakers.

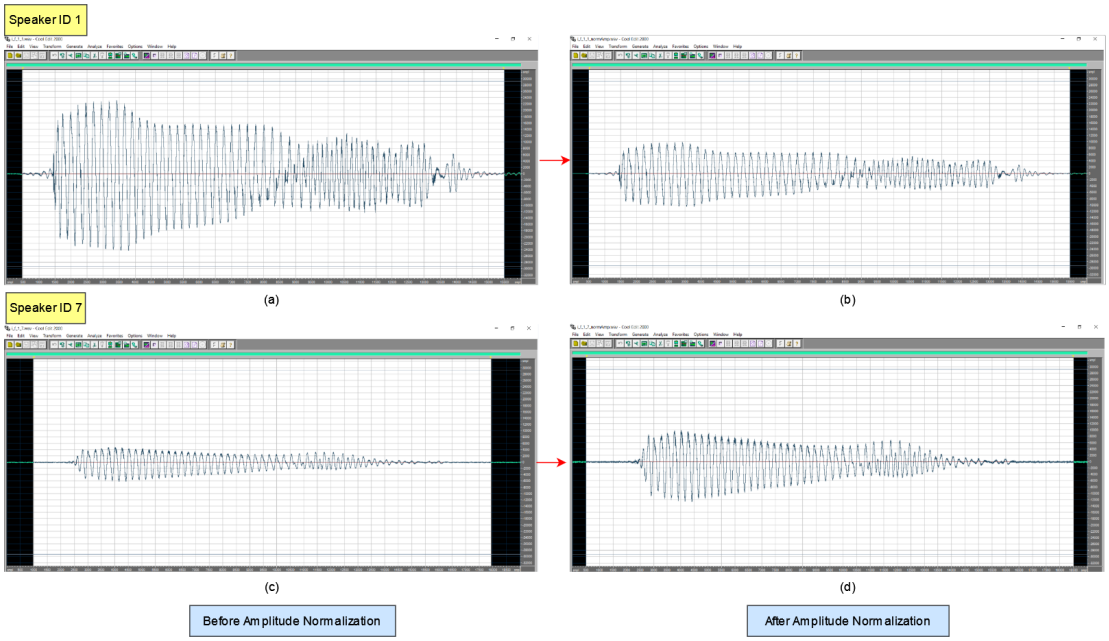


Fig. 5.1: Cooledit interface: waveform of before and after amplitude normalization of 2 speakers' speech utterances.

5.4 Fundamental Frequency and Pitch

The tone is closely related to the fundamental frequency (F_0) and pitch, which is the perceptual interpretation of the fundamental frequency. Fundamental frequency refers to the rate at which our vocal cords vibrate per second when producing voiced sounds. This frequency is quantified in Hertz (Hz), where one Hz signifies one cycle per second. Pitch, on the other hand, represents the perceived F_0, determining whether a signal is perceived as high or low in pitch. Thus, pitch serves as the fundamental element for understanding tone. It is important to emphasize that pitch and fundamental frequency (F_0) are used interchangeably in this context.

Praat (v 6.1.51) [6] is a tool that can analyze, synthesize and manipulate speech data. The speech samples from the developed dataset are analyzed using the Praat software. In tone analysis, features reflecting the pitch contour hold lexical significance. The fundamental frequency, denoted as F_0, indicates tone. In the preliminary study

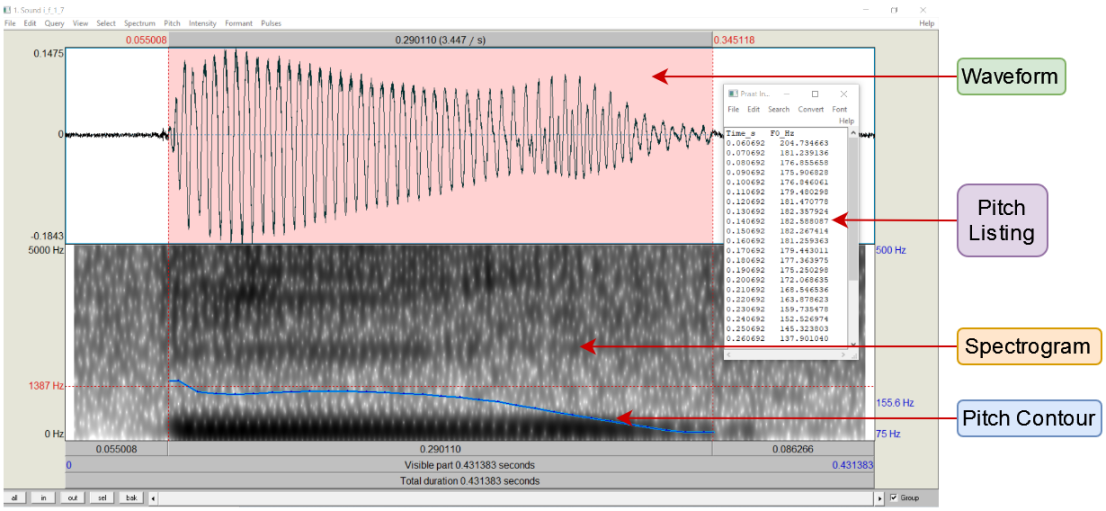


Fig. 5.2: Praat interface.

conducted on the ManiTo dataset, pitch or F0 is extracted using Praat for further analysis and examination.

Figure 5.2 displays a screenshot illustrating the extraction of pitch listing values for a specific WAV file in Praat. The upper portion depicts the waveform of the speech sample, while the lower portion presents the spectrogram. Within the spectrogram, the blue line visually represents the pitch contour. On the right side, a list of numbers indicates the pitch (F_0) listing values in Hertz (Hz) alongside their respective time points.

Pitch is primarily determined by the frequency of the sound waves produced, which corresponds to the rate of vibration of the vocal cords or the source of the sound. Amplitude normalization primarily affects the volume or loudness of the signal by scaling the amplitude values while preserving the relative frequency information. However, it is essential to note that certain forms of amplitude normalization, particularly those involving aggressive compression or equalization, may indirectly affect the perceived pitch by altering the spectral balance of the sound signal. In such cases, careful attention should be paid to the specific methods and parameters used in the normalization process to minimize unintended alterations to the pitch. We ensured that amplitude

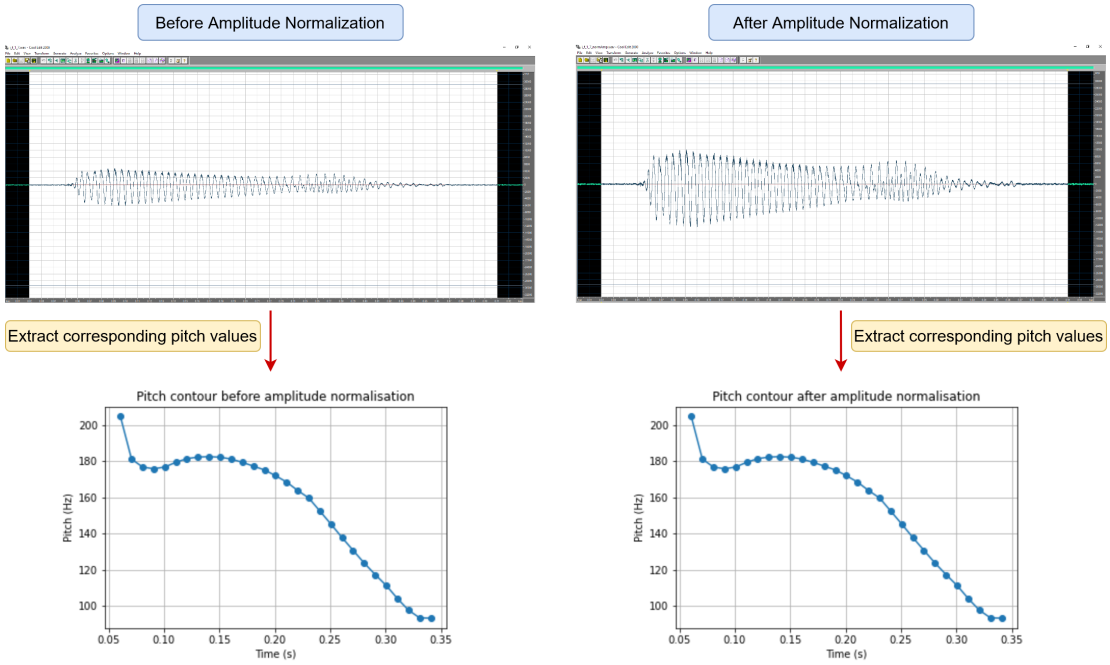


Fig. 5.3: Illustrating absence of pitch contour alteration in post-amplitude normalization.

normalization performed in Section 5.3 does not affect the pitch by plotting the pitch contour corresponding to the speech sample before and after normalization, as shown in Fig. 5.3. By examining the graphs, we can confirm no discernible alteration in the pitch. Hence, it is demonstrated that the amplitude normalization performed in Section 3 does not impact the pitch of the audio signal.

5.5 Analysis of the Pitch Variations in Manipuri Tones

To investigate the tonal variations in Manipuri, explicitly focusing on the **falling tone** and **level tone** distinctions, our initial analysis was conducted with a single speaker and a pair of tonal contrast words in Manipuri: 'un' where the falling tone ꯎꯟ꯭ꯆ/ùn/ signifies 'skin' and the level tone ꯎꯆ/un/ denotes 'snow.' We choose Speaker 1 for this analysis.

We gathered five utterances of the falling tone word and five of the level tone word

from Speaker 1. Subsequently, we extracted the pitch using Praat, equipped with a highly precise pitch extraction algorithm [5]. The pitch values obtained from all ten utterances were collected and stored in separate files.

Next, we generated pitch contours for the five utterances of the falling tone word and the five of the level tone word, as illustrated in Fig. 5.4a and 5.4b, respectively. The graphs show that the pitch contours vary in length and commence at different times across utterances. This disparity underscores the necessity for pitch normalization to ensure comparability and consistency in analyzing the tonal variations.

Algorithm 2: Normalization of length of pitch contour

Input:
$f0[L]$: Pitch listing containing L values
L_{max}: Maximum Pitch listing length
Output:
$Norm_{f0}[L_{max}]$: Pitch listing with L_{max} values

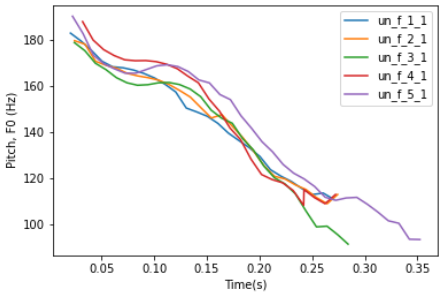
```
1  begin
2      Step 1: k ← 0                                   ▷ f0 index
3      Step 2: j ← 0                                   ▷ track insertion point
4      Step 3: if L < L_max then
5          insert_Loc ← L/ (L_max − L)
6          for i ← 0 to L_max do
7              Norm_f0[i] ← f0[k]
8              if j = insert_Loc then
9                  i ← i + 1
10                 Norm_f0[i] ← f0[k]
11                 j ← 0
12             k ← k + 1
13             j ← j + 1
14     else
15         Norm_f0 ← f0
```

Pitch normalization is essential for tone analysis to guarantee that the pitch contours of various utterances are comparable and consistent, simplifying the interpretation

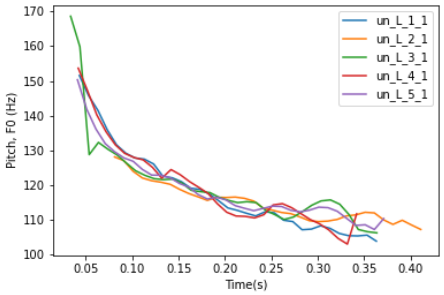
of tonal patterns in speech data. The normalization process involves treating each extracted pitch listing value to enable efficient comparison of the five utterances of each tone type. This normalization ensures that utterances of the same tone type possess uniform starting points and lengths of pitch listing values.

Algorithm 2 outlines normalizing a pitch contour to ensure it fits within a specified maximum length, denoted as L_{max}. L_{max} is the maximum length of pitch listing values among the compared utterances. The algorithm begins by assessing whether the original length (L) of the pitch contour of the utterance to be normalized is less than L_{max}. If equal, the normalized pitch listing, $Norm_{f0}$, remains identical to the original pitch listing values, $f0$. If the original length, L is less than L_{max}, the algorithm calculates the location after which the pitch values need to be inserted to reach the total length of L_{max}, denoted as $insert_{Loc}$. For every $insert_{Loc}$, the algorithm repeats the current value of the original pitch in the $Norm_{f0}$, resulting in a normalized pitch listing with L_{max} values. The algorithm's output is the normalized length pitch listing, $Norm_{f0}$, comprising L_{max} values. Thus, the algorithm normalizes a pitch contour by inserting values from the original pitch listing at regular intervals to achieve a specified maximum length. This process helps standardize pitch contours for consistent analysis and comparison.

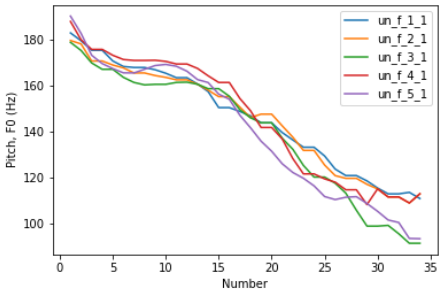
Figure 5.4 comprehensively compares the tonal contrast pair 'un' spoken by Speaker 1. In Fig. 5.4a, the pitch contours of five utterances of the falling tone word are displayed, where each line represents the pitch trajectory of a single utterance. Similarly, Fig. 5.4b showcases the pitch contours of five utterances of the level tone word. Figures 5.4c and 5.4d present the pitch contours of the falling and level tones after pitch normalization. This normalization ensures consistency in pitch lengths across the same tone type utterances, enabling accurate comparison. Furthermore, in Fig. 5.4e, the average pitch listing between the falling tone, represented by the orange line and the level tone, represented by the blue line, is compared. This comparison provides valuable insights



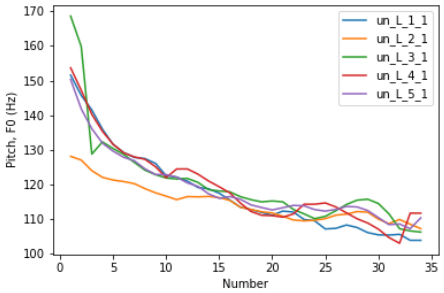
(a) Pitch contour of five utterances of falling tone /ùn/



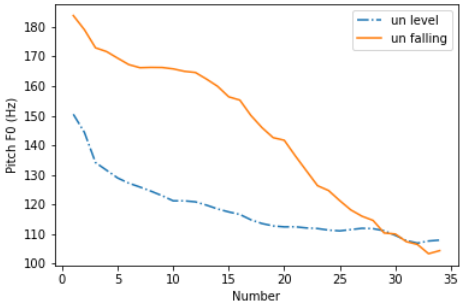
(b) Pitch contour of five utterances of level tone /un/



(c) Normalise pitch contour of falling tone utterances



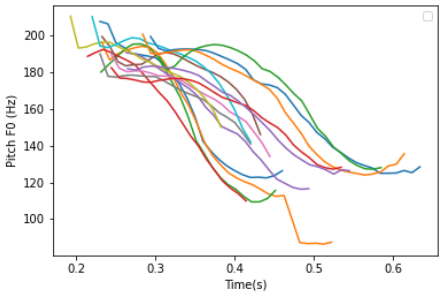
(d) Normalise pitch contour of level tone utterances



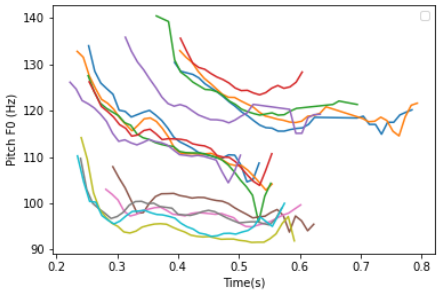
(e) Average pitch contour of five utterances falling and level tones

Fig. 5.4: Pitch contour analysis of Falling and Level tones of Manipuri

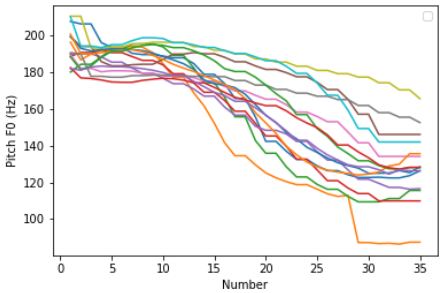
into their respective characteristics and differences. Initial observation reveals that the pitch of the falling tone tends to be higher than that of the level tone. The falling tone typically begins with a high F_0 value and steadily decreases, while the level tone maintains a relatively constant pitch. This graphical representation allows for a detailed analysis of pitch characteristics associated with the two tones of Manipuri, facilitating a deeper understanding of tonal variations and patterns across utterances.



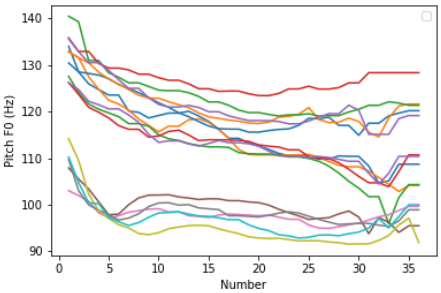
(a) Pitch Contour of 15 Falling Tone Word Utterances for Male Speakers



(b) Pitch Contour of 15 Level Tone Word Utterances for Male Speakers



(c) Normalised Pitch Contour of 15 Falling Tone Word Utterances

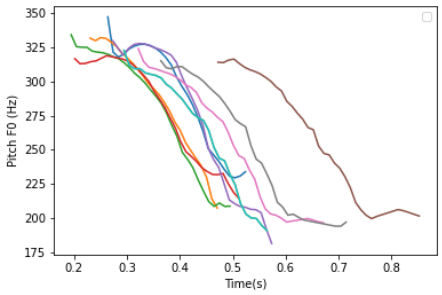


(d) Normalised Pitch Contour of 15 Level Tone Word Utterances

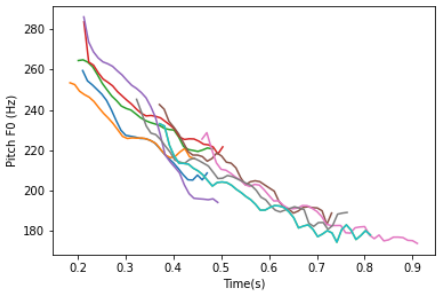
Fig. 5.5: Pitch analysis of Falling and Level Tones in three male speakers for the tonal contrast pair 'sing'

Further pitch analysis is conducted based on gender, involving three male speakers (Speaker ID 1, 3 and 5) and two female speakers (Speaker ID 2 and 4). The tonal contrast word 'sing' is selected, where the falling tone (ꯍꯣꯡ꯫) signifies 'firewood' and the level tone (ꯍꯣꯡ) denotes 'ginger' in Manipuri. Pitch values from 15 utterances of each tone

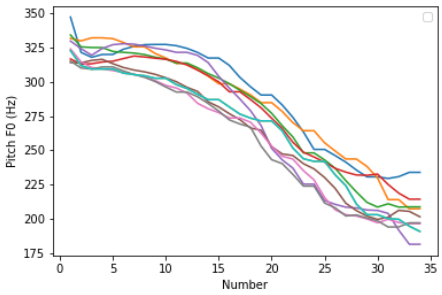
type are extracted for the three male speakers and plotted to visualize the pitch contour. In Fig. 5.5a, the 15 pitch contours of the falling tone utterances for the three male speakers are displayed, with each line representing the pitch contour of an utterance. Similarly, Fig. 5.5b illustrates the pitch contours for the level tone. However, comparison becomes challenging due to variations in starting points and lengths among the utterances. Pitch length normalization (Algorithm 2) is applied to address this. Subsequently, Fig. 5.5c shows the normalized pitch contours of the 15 utterances of the falling tone, while Fig. 5.5d displays those of the level tone for the male speakers.



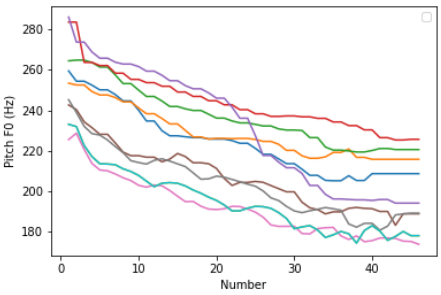
(a) Pitch Contour of 10 Falling Tone Word Utterances for Female Speakers



(b) Pitch Contour of 10 Level Tone Word Utterances for Female Speakers



(c) Normalised Pitch Contour of 10 Falling Tone Word Utterances



(d) Normalised Pitch Contour of 10 Level Tone Word Utterances

Fig. 5.6: Pitch analysis of Falling and Level Tones in two female speakers for the tonal contrast pair 'sing'

Figure 5.7 presents the average pitch contour comparison of the 15 utterances of falling tone and level tone, where the orange line represents the falling tone and the

blue line represents the level tone. It is observed that the falling tone has significantly higher pitch values compared to the level tone and its frequency starts high and falls steadily. In contrast, the level tone maintains consistent frequency.

Similar analyses are conducted for the female speakers, extracting 10 utterances of each tone type for the word 'sing.' Fig. 5.6a and Fig. 5.6b illustrate the pitch contours of the 10 utterances of falling tone and level tone, respectively, with each line representing the pitch contour of an utterance. Pitch length normalization is then performed, resulting in Fig. 5.6c displaying the normalized pitch contours of the falling tone utterances and Fig. 5.6d showing those of the level tone utterances.

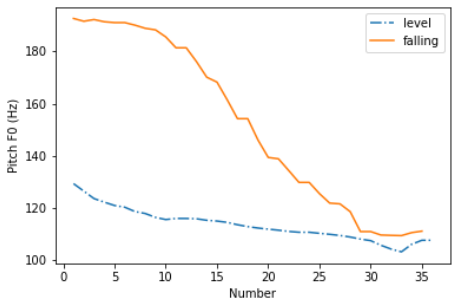


Fig. 5.7: Average pitch contours of 15 utterances of each Falling and Level Tones by male speakers

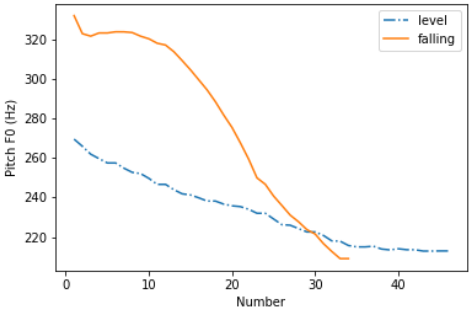


Fig. 5.8: Average pitch contours of 10 utterances of each Falling and Level Tones by female speakers

Finally, Fig. 5.8 presents the average pitch contour of the 10 falling tone utterances (orange) and the average level tone (blue). Figures 5.7 and 5.8 show female voices have higher pitch values than male voices. The average starting pitch for falling tone in females exceeds 320 Hz, while for males, it is below 200 Hz. Similarly, females exhibit pitches above 260 Hz for the level tone, surpassing the starting point of males' falling tone and level tone below 140 Hz. Hence, it is concluded that male and female voices need to be normalized or analyzed separately for tone recognition in Manipuri.

5.6 Conclusion

The initial analysis of the ManiTo, Manipuri Tonal Contrast dataset suggests that pitch contour can effectively distinguish between the two tones, namely falling and level tones. The study indicates that characteristics such as the slope and height of the pitch contour could serve as distinguishing factors for the two tones in Manipuri. Furthermore, the analysis reveals that pitch characteristics in male and female speech data are not directly comparable, indicating the necessity to analyze speech samples separately based on gender. Consequently, the study underscores the importance of gender speech normalization in future analyses to ensure accurate and meaningful comparisons between male and female speech data in Manipuri tonal contrast analysis. Moreover, similar studies on other tonal contrast words spoken by different speakers revealed consistent characteristics across the dataset, further reinforcing the reliability and generalization of pitch contour to differentiate between the two tones in Manipuri.

6

Disambiguation of Isolated Manipuri Tonal Contrast Word Pairs using Acoustic Features

Manipuri is a low-resource, Tibeto-Burman tonal language spoken mainly in Manipur, a northeastern state of India. Tone identification is crucial to speech comprehension for tonal languages, where tone defines the word's meaning. Automatic Speech Recognition for those languages can perform better by including tonal information from a powerful tone detection system. While significant research has been conducted on tonal languages like Mandarin, Thai, Cantonese and Vietnamese, a notable gap exists in exploring Manipuri within this context. To address this gap, this study used our previously developed handcrafted speech corpus, ManiTo, which comprises isolated Manipuri tonal contrast word pairs to study the tones of Manipuri. The corpus consists of 17,837 labeled samples from twenty native speakers. Preliminary findings have confirmed that Manipuri has two unique tones, Falling and Level. A comprehensive acoustic feature analysis is conducted next. Two sets of features based on Pitch contours, jitter and shimmer measurements are investigated to distinguish the two tones of Manipuri. Support Vector Machine, Long Short-Term Memory, Random Forest and k-Nearest Neighbors are the classifiers adopted to validate the selected feature sets. The results indicate that the second set of features consistently outperformed the first set, demonstrating higher ac-

curacy, particularly when utilizing the Random Forest classifier. This provides valuable insights for further advancements in speech recognition technology for low-resource tonal language Manipuri.

6.1 Related Work

Extensive research has been done in the last few decades on the tone recognition of tonal languages, such as Mandarin, Thai, Cantonese and Vietnamese. Few works have been reported on the Indian tonal language Mizo, Bodo, Punjabi and the African continental tone languages such as Yoruba and Hausa [73].

Tone recognition of Chinese syllables was proposed based on the Support Vector Machine (SVM) and Deep Belief Network (DBN) model [11]. The DBN was used to extract the prosodic and articulatory features and SVM was used for the tone classification. The method was applied on the 863-data corpus with an average accuracy rate achieved of 83.03%. A multi-scale model was proposed to capture the features of tone variation of Mandarin affected by complex linguistic and phonetic rules [63]. Cepstrogram and Mel-spectrogram were used as the input features of the speech data and a multi-scale Convolutional Neural Network (CNN) was employed to recognize the tone. Experimental analysis showed the method achieved state-of-the-art results on the Corpus 863 data.

Yan et al. [92] proposed a method for **Mandarin tone identification** based on Random Forest (RF) and three feature fusion sets (FFSs). The Syllable Corpus of Standard Chinese (SCSC), a speaker-independent Mandarin monosyllable corpus, was used to assess the methodology. The findings demonstrate that, even with unbalanced data, the tone recognition algorithm can achieve high tone recognition accuracy and exhibit strong generalization and classification capabilities. The SCSC dataset is again used to evaluate

the model designed using CNN named ToneNet [25], which applies Mel-spectrogram as features, achieving a tone recognition rate of 99.16% accuracy.

Table 6.1: Some work done on tone recognition of tonal languages

Author and Year	Language	Dataset	Features	Classifier	Accuracy(%)
Chao et al. 2019 [11]	Chinese	863 corpus	Prosodic & articulatory extracted by DBN	SVM	83.03
Gao et al. 2019 [25]	Mandarin	SCSC	Mel-spectrogram	ToneNet: CNN Network	99.16
Peng et al. 2021 [63]	Mandarin	Hi-Tech Project 863 corpus	Mel-spectrogram cepstrogram	CNN	88.06
Yan et al. 2023 [92]	Mandarin	SCSC	FFSs	RF	93.57
Potisuk et al. 1999 [64]	Thai	5 speakers of total 625 utterances	Normalized F_0 contour	Extended Fujisaki model	89.10
Kertkeidkachorn et al. 2015 [42]	Thai	12 speakers (1320 samples), LOTUS	Energy based and spectral based	HCRF	94.24 71.01
Sarma et al. 2015 [68]	Mizo	Collected from 5 Mizo Speakers	Pitch contour	Based on variance, height and slope of pitch	70.00
Gogoi et al. 2020 [27]	Mizo	Collected from 19 native speakers	F_0 contour	SVM DNN	73.39 74.11

In Figure 6.4, the tone geometrical model was presented to derive the seven selected features for the five Thai tones based on pitch contour [14]. Experimental analysis showed that the selected parameters can distinguish the five tones of the Thai language. Potisuk et al. [64] proposed a **Thai tone recognition** system using an extension of Fujisaki's model [24] that incorporates tonal declination and assimilation, achieving an accuracy rate of 89.1%. Another study in Thai tone recognition was presented by Kertkeidkachorn et al. [42]. Hidden Conditional Random Field (HCRF) was used as a tone classifier, applying energy-based and spectral-based features, including Linear Prediction Coefficients (LPC), Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). In isolated words, energy-based systems provide an error rate reduction of 22.40%, and in the continuous-speech, PLP-based features achieved an error rate reduction of 13.90%.

Mizo tone recognition was proposed using features derived from pitch contour

and applied on SVM based and Deep Neural Network (DNN) based models achieving an accuracy of 73.39% and 74.11% respectively for the two models [27]. Another tone detection of Mizo is conducted using variance, height and slope of pitch values where the recognition accuracy was 70% [68]. Table 6.1 shows the summarised comparison of work based on tone recognition.

On the other hand, work has yet to be conducted in the field of **tone recognition for Manipuri**. Despite the significance of tone in the Manipuri language, research efforts addressing tone recognition have not been addressed adequately. This gap in the literature highlights the need for dedicated studies to develop tone recognition systems tailored to Manipuri, which could greatly benefit language learning, linguistic research and culture preservation efforts.

6.2 Materials and Techniques

The subsequent subsections delve into three crucial areas: The development of the dataset for Manipuri Tonal Contrast words, the methodology employed for feature extraction and the selection of classifiers. This comprehensive examination aims to provide a detailed exploration of each fundamental aspect, contributing to a thorough understanding of the research methodology.

6.2.1 Dataset for Manipuri Tonal Contrast Words

The developed speech corpus ManiTo in Chapter 4 is a comprehensive isolated tonal contrast speech database meticulously designed to investigate the two tones, level and falling, inherent in the Manipuri language. The database contains 54 tonal contrast word pairs gathered from 20 native speakers, comprising 11 males and 9 females aged between 19 and 70. Recordings are captured with a 16-bit resolution and a sampling rate

of 44,100 Hz, utilizing a mono channel through Cool Edit 2000 (v1) [79]. Each recorded speech segment undergoes meticulous analysis, manual segmentation and labeling, including word name, tone type, utterance number and speaker ID. Presently, the dataset encompasses 17,837 samples of labeled speech data, offering a valuable resource for further research and analysis into Manipuri tonal contrasts and associated linguistic phenomena.

6.2.2 Feature Extraction

A sound wave's fundamental frequency (F_0) correlates to the perceptual characteristic of the sound called pitch. The pitch has been regarded as the vital cue in tone recognition of tonal languages [27]. Initially, pitch contour analysis is done for a pair of Manipuri tonal contrast words using Praat (v 6.1.51) [6].

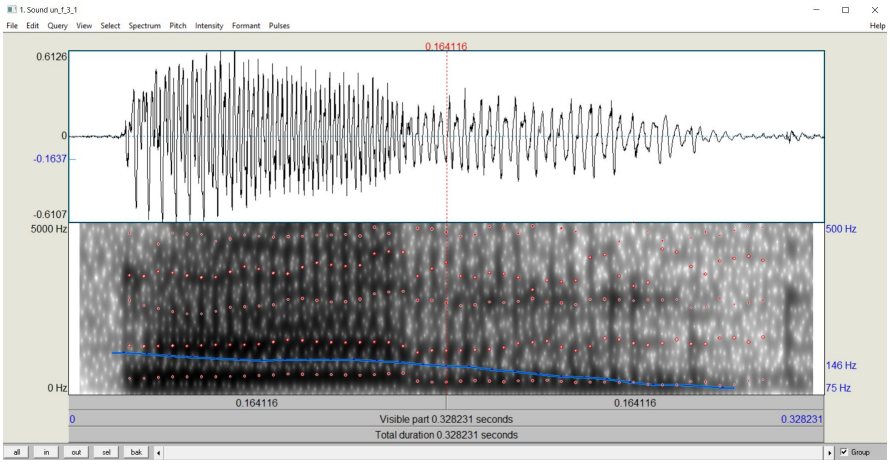


Fig. 6.1: Pitch contour and waveform of falling tone /ùn/ (ꯎ꯭ꯆ) in Praat.

Figure 6.1 and Figure 6.2 show the waveform and pitch contour (blue line) of the falling tone word /ùn/ (ꯎ꯭ꯆ) meaning skin and level tone word /un/ (ꯎꯆ) meaning snow respectively. In Figure 6.1, the falling tone has its pitch onset at 146 Hz and falls steadily for about 71 Hz, while in Figure 6.2, the level tone pitch stays relatively constant.

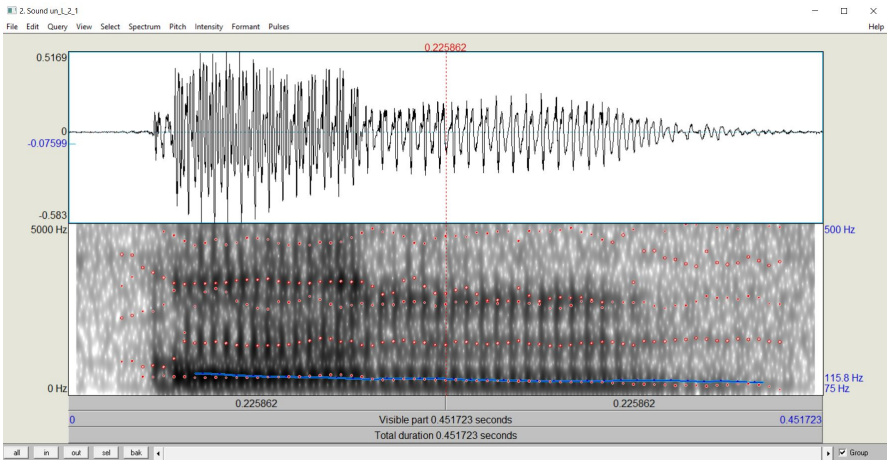


Fig. 6.2: Pitch contour and waveform of level tone /*un*/ (ꯎꯅ) in Praat.

Further analysis is done on the other tonal contrast word pairs and it is observed that similar characteristics are found in the respective tone types. Each speech sample amplitude is brought to a particular volume level by applying the peak normalization technique outlined in Algorithm 1 of Chapter 5 so that the speech samples are consistent across the different environments, recordings or speakers as varying loudness levels make it difficult to compare or analyze the speech data across samples.

The Python package Parselmouth [39] extracts the F_0 contour of speech samples. Figure 6.3 shows the averaged F_0 contour of 15 utterances of falling tone word /sìŋ/ (ꯁꯤꯡ꯫) meaning firewood indicated by the orange line and 15 utterances of level tone word /siŋ/ (ꯁꯤꯡ) meaning ginger indicated by the blue line. The orange line has a steady slope as the frequency contained in the falling tone starts high and decreases rapidly with time, whereas the blue line stays relatively at one level.

Disambiguation of the two tones of Manipuri highly depends on the F_0 contour. The following five parameters are derived from the pitch contour [14] of the speech samples to detect the falling and the level tone:

- **Pitch Duration:** In Figure 6.4, pitch duration (T) is the length of the pitch contour,

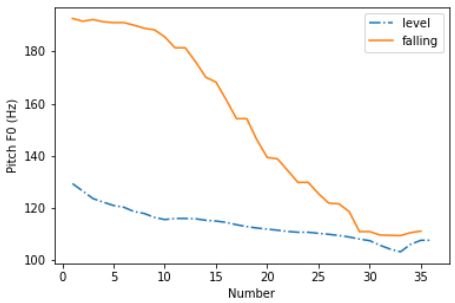


Fig. 6.3: Average F_0 contour for the falling and level tones of Manipuri

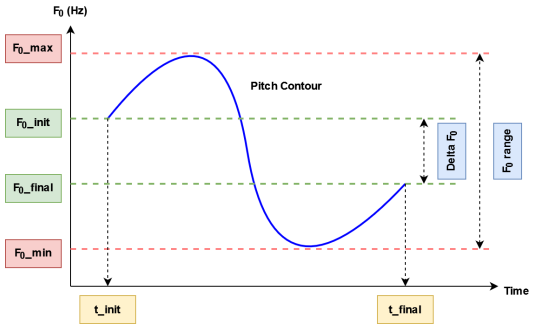


Fig. 6.4: Geometrical tone model extracted from pitch contour [14]

given by (6.1)

$$T = t\_final - t\_init \tag{6.1}$$

where,

t_final is the pitch contour's final time point

t_init is the pitch contour's initial time point

- **Initial F_0:** The starting value of the pitch contour, denoted by F_0_init in Figure 6.4.

- **Delta F_0 (δF_0):** It represent the frequency variation between the initial F_0 (F_0_init) and final F_0 (F_0_final), given by (6.2)

$$\delta F_0 = F_0\_final - F_0\_init \tag{6.2}$$

- **Sign F_0 range:** The frequency variation between the maximum value of F_0 (F_0_max) and minimum value of F_0 (F_0_min) is represented by F_0 range (F_0_ran). Sign of F_0 range ($sign\_F_0\_ran$) indicated the direction of the F_0_ran where negative

means downward and positive means upward movement, given by (6.3)

$$sign\_F_0\_ran = sign\_\delta F_0 * F_0\_ran \tag{6.3}$$

where,

$$sign\_\delta F_0 = \begin{cases} 1 & \text{if } \delta F_0 \text{ is positive} \\ -1 & \text{if } \delta F_0 \text{ is negative} \end{cases}$$

and

$$F_0\_ran = F_0\_max - F_0\_min$$

- **Sign F_0 slope ($sign\_F_0\_slope$):** It is the ratio of sign F_0 range to pitch duration, given by (6.4)

$$sign\_F_0\_slope = sign\_F_0\_ran/T \tag{6.4}$$

In the preliminary analysis of the two tones, the above five pitch features were studied on a tonal contrast word pair /càbə/ (ञ्रे·ठ), which means eating, bearing falling tone and level tone word /cabə/ (ञ्रेठ) meaning swim visibly. Table 6.2 represents the mean and standard deviation (SD) values for the five features extracted from 35 utterances of each tone type. It is observed that the parameters of the two tones have different magnitudes of mean and SD values. The parameters of the falling tone are comparatively higher than the level tone.

Table 6.2: Statistical values of pitch derived parameters for the falling tone and level tone of a word pair 'caba'

	Tone Type	Pitch Duration (second)	Initial F_0 (Hz)	Delta F_0 (Hz)	Sign F_0 Range (Hz)	Sign F_0 Slope (Hz)
Mean	Falling	0.619	271.727	-96.350	-111.070	-180.752
	Level	0.532	204.749	-12.382	-41.905	-79.750
Standard Deviation	Falling	0.056	9.355	42.215	37.964	62.644
	Level	0.081	7.547	22.985	43.248	78.612

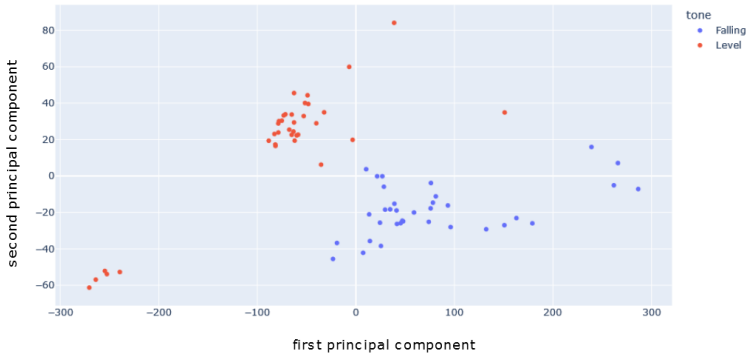


Fig. 6.5: Projection of the 5 pitch feature on 2D

Figure 6.5 shows the extracted features of each speech sample projected on two-dimensional space using Principal Component Analysis (PCA) [81]. Each blue dot represents the five pitch-derived features of a falling tone speech projection and each red dot represents a level tone speech's features on the 2-D space. It is noticed that there is a clear separation of the two tones of Manipuri, a hyperplane can segregate them.

Further features investigation is performed based on the jitter and shimmer of the speech sample. Jitter and shimmer represent variations in vocal signals' fundamental frequency F_0. While shimmer emphasizes the amplitude or intensity of vocal emission, jitter pertains to the frequency fluctuation of the signal's F_0 [86]. These measurements of F_0 disturbance, jitter and shimmer are valuable in describing vocal characteristics, with jitter reflecting the parameter of frequency variation from cycle to cycle and shimmer indicating the amplitude variation of the sound wave [83]. There are various forms of jitter and shimmer. The different measurements of jitter and their computational formulas [83, 6, 21] are as follows:

- **Local Absolute Jitter ($local_{absolute}$):** Average absolute difference between succes-

sive periods, measured in seconds.

$$local_{absolute} = \frac{1}{n-1}\sum_{i=1}^{n-1}|T_i - T_{i-1}| \tag{6.5}$$

where n is the total number of intervals and T_i is the time period of the i^{th} interval.

- **Local Jitter ($local$, $relative$):** It is calculated by dividing the $local\ absolute$ Jitter by the average time period.

$$local = \frac{local_{absolute}}{\frac{1}{n}\sum_{i=1}^{n} T_i} \times 100 \tag{6.6}$$

- **Relative Average Perturbation Jitter (rap):** Represent the disturbance, that is, the average absolute difference between a time period and the average of it and its two neighbors, divided by the average time period.

$$rap = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}|T_i - \left(\frac{1}{3}\sum_{m=i-1}^{i+1} T_m\right)|}{\frac{1}{n}\sum_{i=1}^{n} T_i} \times 100 \tag{6.7}$$

- **5 Point Period Perturbation Quotient Jitter ($ppq5$):** Represent the ratio of disturbance that is the absolute difference between a time period and the average of it and its four closest neighbors (two prior and two successive periods), divided by the average time period.

$$ppq5 = \frac{\frac{1}{n-1}\sum_{i=2}^{n-2}|T_i - \left(\frac{1}{5}\sum_{m=i-2}^{i+2} T_m\right)|}{\frac{1}{n}\sum_{i=1}^{n} T_i} \times 100 \tag{6.8}$$

- **DDP Jitter (ddp):** Represent the ratio of the average absolute difference between consecutive differences between consecutive time periods and the average time period.

$$ddp = \frac{\frac{1}{n-2}\sum_{i=2}^{n-1}|(T_{i+1}-T_i)-(T_i-T_{i-1})|}{\frac{1}{n}\sum_{i=1}^{n}T_i} \times 100 \tag{6.9}$$

The different measurements of shimmer and their computational formulas [83, 6, 21, 20] are as follows:

- **Shimmer (dB):** Expresses the variability in the peak-to-peak amplitude, measured in decibels. It is computed as the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods multiplied by 20.

$$dB = \frac{1}{n-1}\sum_{i=1}^{n-1}|20 * log(\frac{a_{i+1}}{a_i})| \tag{6.10}$$

  where a_i represent the extracted peak-to-peak amplitude data and n denotes the number of extracted fundamental frequency periods.

- **Shimmer ($local, relative$):** Indicates the average absolute difference between the amplitudes of two successive periods, divided by the mean amplitude.

$$relative = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}|a_i - a_{i+1}|}{\frac{1}{n}\sum_{i=1}^{n}a_i} \times 100 \tag{6.11}$$

- **Three-point Amplitude Perturbation Quotient Shimmer ($apq3$):** Represents the quotient of amplitude disturbance within three periods. It is the average absolute difference between the amplitude of a period and the mean amplitudes of its two neighboring periods divided by the average amplitude.

$$apq3 = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}|a_i - \left(\frac{1}{3}\sum_{j=i-1}^{i+1}a_j\right)|}{\frac{1}{n}\sum_{i=1}^{n}a_i} \times 100 \tag{6.12}$$

- **Five-point Amplitude Perturbation Quotient Shimmer ($apq5$):** Represents the

ratio of perturbation amplitude observed over five consecutive periods. It calculates the average absolute difference between the amplitude of a period and the mean amplitudes of that period and its four nearest neighbors. This value is then divided by the average amplitude to derive the final ratio.

$$apq5 = \frac{\frac{1}{n-1}\sum_{i=2}^{n-2}|a_i - \left(\frac{1}{5}\sum_{j=i-2}^{i+2} a_j\right)|}{\frac{1}{n}\sum_{i=1}^{n} a_i} \times 100 \tag{6.13}$$

- **Eleven-point Amplitude Perturbation Quotient Shimmer ($apq11$):** Represents the 11-point Amplitude Perturbation Quotient, which quantifies the average absolute difference between the amplitude of a period and the average of the amplitudes of that period and its ten closest neighbors. This value is then divided by the average amplitude of the speech signal.

$$apq11 = \frac{\frac{1}{n-10}\sum_{i=6}^{n-5}|a_i - \left(\frac{1}{11}\sum_{k=i-5}^{i+5} a_k\right)|}{\frac{1}{n}\sum_{i=1}^{n} a_i} \times 100 \tag{6.14}$$

Table 6.3: Jitter and shimmer measurements for a tonal contrast pair

	Measurement	Utterance	
		un_f_1_1 (falling tone)	un_L_1_1 (level tone)
Jitter	local, absolute	127.292E-6 s	117.728E-6 s
	local (%)	1.820	1.402
	rap(%)	0.678	0.497
	ppq5(%)	0.794	0.759
	ddp(%)	2.034	1.491
Shimmer	dB	0.771 dB	0.723 dB
	local(%)	7.944	7.657
	apq3(%)	2.988	2.753
	apq5(%)	5.012	4.011
	apq11(%)	8.264	8.047

The analysis of jitter and shimmer was conducted on a pair of tonal contrast words using the sample 'un_f_1_1' representing the falling tone word (ඝුං), uttered by SpeakerID 1 and 'un_L_1_1' representing the level tone word (ඝුං), spoken by SpeakerID 1

as well. Using the Praat tool, the measurements of jitter and shimmer were initially extracted. Table 6.3 presents the values for the ten measurements of the selected falling and level tone words. It is observed that there are slight differences in the values, with the falling tone showing slightly higher values compared to the level tone. A similar analysis is done across other pairs and speakers, revealing similar characteristics.

The five different jitter and shimmer measurements are extracted for each speech sample in ManiTo using the Parselmouth package [39]. PCA is applied to the extracted jitter and shimmer features. PCA [71] transforms the original features into a new set of features, called principal components, which capture the maximum variance in the data. By doing so, PCA reduces the dimensionality of the data while retaining most of its essential information. It removes unnecessary dimensions and includes only the most critical measurements.

In this work, we have proposed two feature sets utilizing PCA:

- **Set1** consists of resultant values after applying PCA on the above 5 extracted pitch derived features.

$$i.e.\ \text{PCA}(T, F_0\_init, \delta F_0, sign\_F_0\_ran, sign\_F_0\_slope)$$

- **Set2** consists of values of Set1, PCA resultant of 5 jitter and 5 shimmer measurements.

$$i.e.\ \text{Set1, PCA}(absolute, local, rap, ppq5, ddp),$$
$$\text{PCA}(dB, local, apq3, apq5, apq11)$$

6.2.3 Classifiers

The objective of the proposed method is to utilize simple computations, support rapid model creation and manage small sample sets efficiently. As a result, the tone recognition classifier's design places a high value on low-resource language and encourages

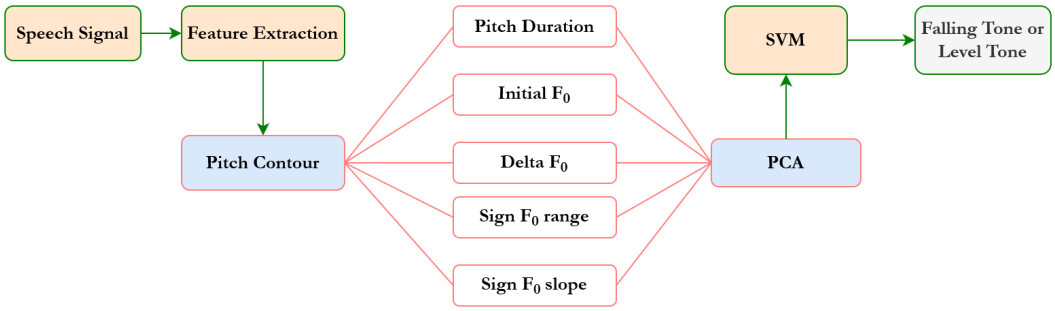


Fig. 6.6: Block diagram of feature Set1 based SVM model for Manipuri tone recognition system

using machine learning models with low complexity. This methodology guarantees applicability and efficacy in execution, conforming to the restrictions of small-scale datasets and resource constraints. Support Vector Machine (SVM) [35], Long Short-Term Memory (LSTM) [36], k-Nearest Neighbors (KNN) [30] and Random Forest (RF) [22] are generally used machine learning classifier in the speech related work [11, 92, 27, 60, 65]. The hyperparameter tuning of the RF and KNN classifier is done using the GridSearch method [72]. In our work, these four classifiers are used to analyze Manipuri's tone recognition system, employing the two proposed feature sets, Set1 and Set2.

6.3 Experimental Analysis and Results

For the preliminary experiment, SVM is adopted to classify the two tones of Manipuri. Figure 6.6 presents the block diagram of the model employing the feature Set1. The five pitch derived parameters are used as the feature vector. PCA [77] is applied to the extracted pitch parameters for feature optimization and the resultant feature is passed to the SVM for training. The performance is evaluated using 5-fold cross-validation [69]. Initially tested on one pair of tonal contrast words, 'caba,' spoken by a single speaker, a total of 70 utterances. The average tone recognition rate for the 5-fold turns out to be 98.57%.

Table 6.4: Performance of the SVM tone recognition system of Manipuri on single male utterances

Data	Feature			
	Set1		Set2	
	Mean Accuracy(%)	SD	Mean Accuracy(%)	SD
Monosyllabic word pairs	92.20	6.73	93.29	5.24
Disyllabic word pairs	91.15	6.12	94.45	3.90
All word pairs	88.40	4.83	92.02	3.11

In our experiments, the k-fold cross-validation technique [69] is employed to evaluate the model's performance to avoid overfitting. By dividing the data into subsets and using different subsets for testing and training, cross-validation helps to provide a more accurate estimate of the model's performance on new data. The value of k (the number of subsets or folds) is assigned to 10 as it gives low bias with good variance [9]. With k = 10, the dataset will be randomly divided into ten parts, nine groups are used for training and one is set aside for testing. This procedure is repeated ten times, saving a different subgroup for testing each time. The mean accuracy and the standard deviation (SD) of 10-fold are used as evaluation matrices of the tone recognition systems.

A pilot experiment with a single male speaker speech samples was conducted to check the mean accuracy and the SD of 10-fold cross-validation using the features Set1 and Set2. Table 6.4 presented the model's performance, trained and tested on three types of data: monosyllabic, disyllabic and all pairs of words in the ManiTo corpus. It is observed that the model achieved 93.29% of the highest accuracy on monosyllabic data employing the feature Set2. On average, there is a 2.67% improvement when the feature Set2 is applied.

A detailed experiment is performed using 20 speakers of 17,837 utterances, of which 11 are males and 9 are females. The first analysis is based on gender and syllable number (monosyllable and disyllable). People have a wide range of vocal pitch. As the feature vectors depend on the pitch, gender-dependent analysis is performed as male voices

Table 6.5: Tone classification results for monosyllabic word pairs

Gender	Feature Set	Result	Classifier			
			SVM	LSTM	RF	KNN
Male	Set1	Mean Accuracy(%)	80.85	80.47	77.73	80.74
		SD	1.96	1.92	1.39	2.47
	Set2	Mean Accuracy(%)	82.51	**82.95**	82.69	81.75
		SD	1.96	1.64	2.09	1.72
Female	Set1	Mean Accuracy(%)	67.50	66.82	69.80	70.63
		SD	2.06	2.11	1.72	2.22
	Set2	Mean Accuracy(%)	74.96	74.71	**77.54**	77.02
		SD	2.40	2.25	2.23	1.82

Note: The highest mean accuracy is highlighted in bold for each gender.

Table 6.6: Tone classification results for disyllabic word pairs

Gender	Feature Set	Result	Classifier			
			SVM	LSTM	RF	KNN
Male	Set1	Mean Accuracy(%)	75.76	75.53	74.50	75.87
		SD	2.21	1.88	0.008	2.24
	Set2	Mean Accuracy(%)	82.34	82.74	**83.83**	82.02
		SD	1.22	1.36	0.01	1.16
Female	Set1	Mean Accuracy(%)	67.81	67.93	68.19	70.76
		SD	2.89	2.65	2.39	1.94
	Set2	Mean Accuracy(%)	77.74	77.47	**79.04**	78.17
		SD	1.53	1.62	1.74	1.34

tend to have a lower pitch than female voices. A comparative experiment on the four classifiers, SVM, LSTM, RNN and KNN, utilizing the two feature sets, Set1 and Set2, is conducted to investigate the two tones of Manipuri. The result in Table 6.5, Table 6.6 and Table 6.7 show the four classifier approach with the tone features, Set1 and Set2 for the monosyllabic word, disyllabic word and all pair of words respectively.

In each experiment, the tone feature Set2 outperforms Set1, achieving the highest mean accuracy of 82.95% and 77.54% for male and female speakers, respectively, for monosyllabic words, 83.83% and 79.04% for disyllabic words and 80.26% and 76.51% for all pairs of words for male and female speaker. Male speech has a higher tone recognition rate than female speech. The distribution of muscle and tissue in the throat and mouth

Table 6.7: Tone classification results for all word pairs

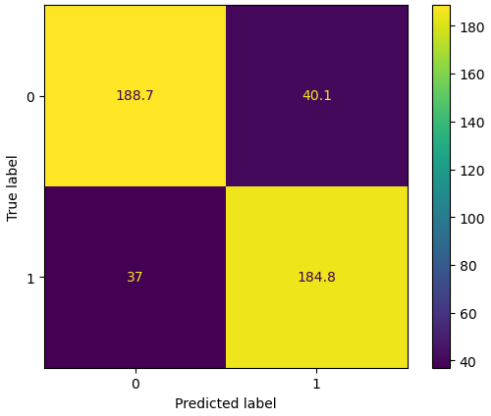
Gender	Feature Set	Result	Classifier			
			SVM	LSTM	RF	KNN
Male	Set1	Mean Accuracy(%)	75.52	75.38	72.65	75.65
		SD	1.04	0.76	1.46	1.003
	Set2	Mean Accuracy(%)	79.53	80.03	**80.26**	79.65
		SD	1.24	0.97	0.84	1.37
Female	Set1	Mean Accuracy(%)	66.08	65.66	65.34	66.97
		SD	1.93	2.01	0.18	1.58
	Set2	Mean Accuracy(%)	73.93	74.47	**76.51**	74.96
		SD	1.69	0.008	1.16	1.00

Table 6.8: Precision, Recall, and F1 Score for Highest Mean Accuracy Using Set2 across Different Data Sets and Genders

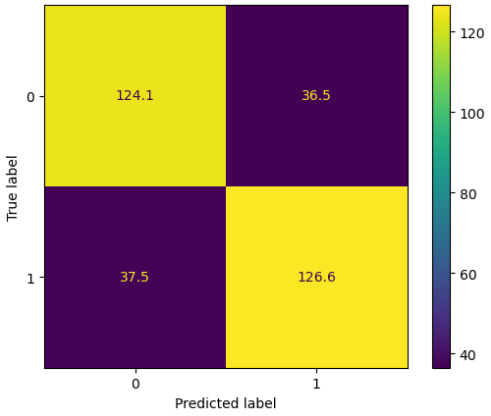
Data	Gender	Model	Mean Accuracy(%)	Precision	Recall	F1 score
Monosyllabic word pairs	Male	LSTM	82.95	82.33	83.33	82.73
	Female	RF	77.54	77.61	77.17	77.35
Disyllabic word pairs	Male	RF	83.83	85.12	82.34	83.67
	Female	RF	79.04	78.94	81.12	79.98
All word pairs	Male	RF	80.26	80.92	79.17	80.01
	Female	RF	76.51	76.57	76.35	76.44

can also affect the pitch and tone of a person's voice. Women tend to have more muscle and tissue in the pharynx and mouth, which can help shape and modify the sound of the voice, resulting in a more complex and varied range of tones and inflections. The RF classifier outperforms the other classifiers on average due to its ensemble nature, flexibility and ability to handle complex datasets while minimizing overfitting, achieving the highest mean accuracy of 83.83%

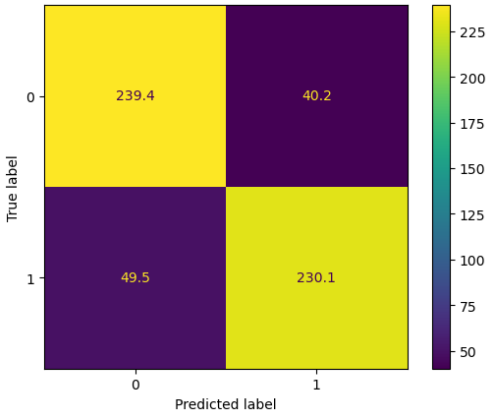
Table 6.8 validates the highest mean accuracy result in Table 6.5, Table 6.6 and Table 6.7 obtained across different word pairs and genders employing the feature Set2 based on metrics such as precision, recall and F1 score. Fig. 6.7 displays the confusion matrix resulting from applying 10-fold cross-validation to the data presented in Table 6.8. The study aims to evaluate how well these models can classify word pairs based on their syllabic structure, considering potential variations across genders.



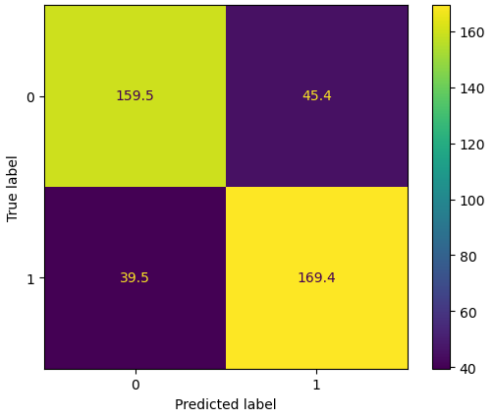
(a) Male monosyllabic word pairs



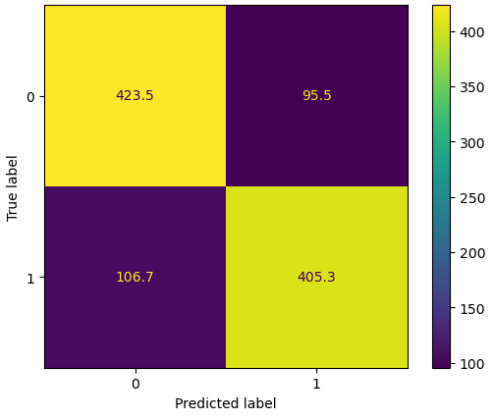
(b) Female monosyllabic word pairs



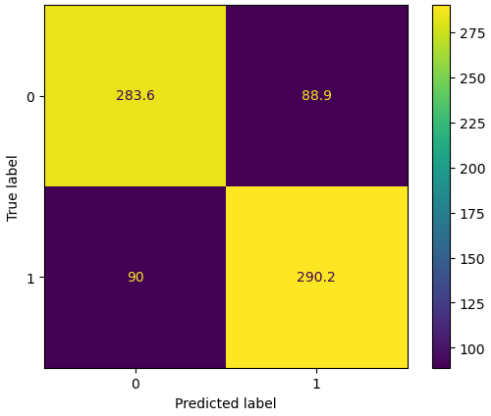
(c) Male disyllabic word pairs



(d) Female disyllabic word pairs



(e) Male all word pairs



(f) Female all word pairs

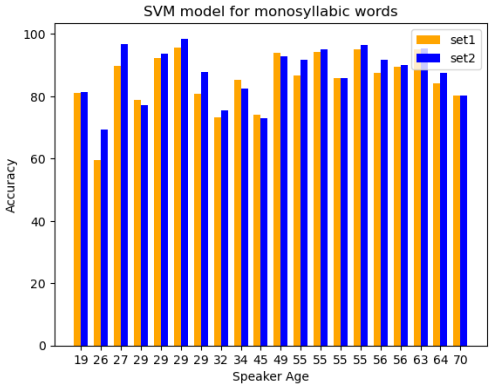
Fig. 6.7: Confusion matrix for different data in Table 6.8

The results indicate varying performance across different types of word pairs and genders. Generally, RF consistently outperformed LSTM across all categories in terms of mean accuracy, precision, recall, and F1 score. Both LSTM and RF achieved relatively high accuracy for monosyllabic word pairs, with RF performing slightly better for both genders. In contrast, for disyllabic and all word pairs, RF consistently demonstrated higher accuracy, precision, recall, and F1 score than LSTM, suggesting RF's robustness in handling different data types. Based on the analysis, Random Forest emerges as a preferable model for classifying word pairs based on syllabic structure, exhibiting superior accuracy and predictive metrics performance across all evaluated categories.

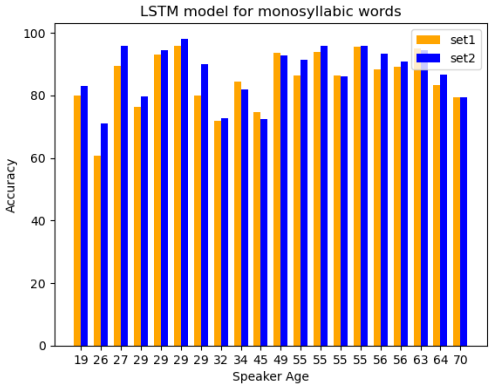
The second analysis focuses on a speaker-dependent approach. The tone features from Set1 and Set2 are employed and individually assessed by the four classifiers, each applied to every speaker separately. The results of tone recognition are depicted in Fig. 6.8 for SVM and LSTM classifiers and in Fig. 6.9 for RF and KNN classifiers, employing feature Set1 and Set2. The evaluation covers various data types for each speaker, including monosyllabic, disyllabic and all word pairs. The bar graph in Fig. 6.8 and 6.9 illustrates tone recognition's 10-fold cross-validation mean accuracy. Feature Set1 is represented by orange, while Set2 is represented by blue.

It is observed that the blue bar is consistently superior to the orange bar in most cases, indicating the better performance of Set2 compared to Set1. Specifically, in KNN classifier Fig. 6.9b, Fig. 6.9d and Fig. 6.9f, for speakers aged 29, 32, 34 and 45 (corresponding to IDs 4, 2, 1 and 3), feature Set1 outperforms Set2. This slight superiority of Set1 is due to the comparatively lower quantity of speech samples for these specific speakers, as detailed in Table 4.3. Despite this observation, the overall performance of KNN is better in the case of Set2 when compared with Set1.

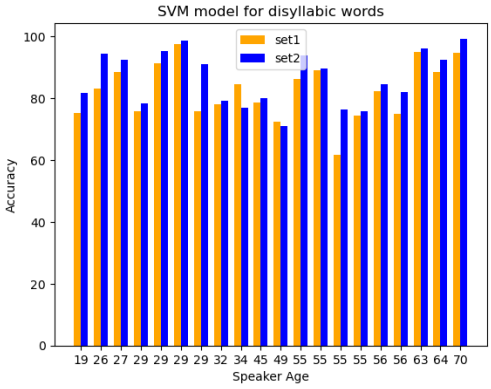
Similarly, Set2 excels in other cases, achieving the highest accuracy of 98.33% (Speaker ID 8, age 29, Fig. 6.9a) for monosyllables using the RF classifier, 99.28% (Speaker ID 20,



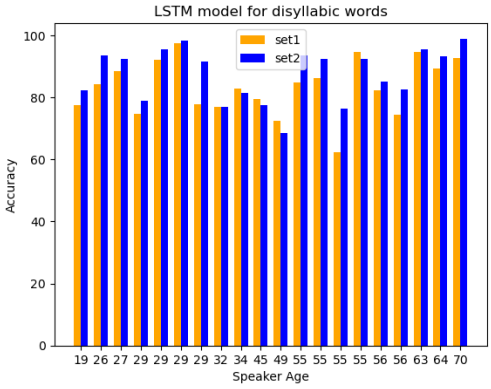
(a) Result of SVM model on monosyllabic words



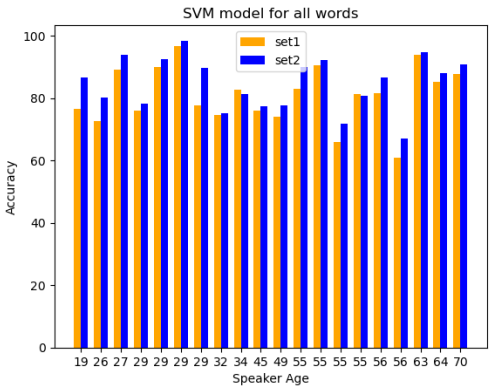
(b) Result of LSTM model on monosyllabic words



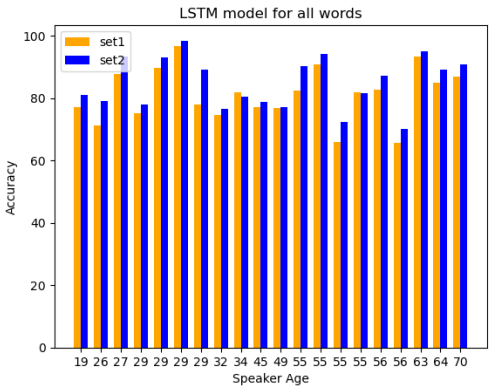
(c) Result of SVM model on disyllabic words



(d) Result of LSTM model on disyllabic words

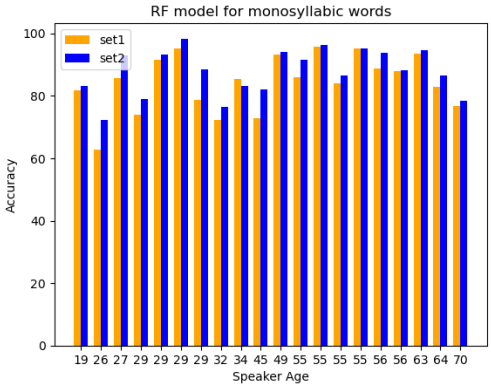


(e) Result of SVM model on all word pairs

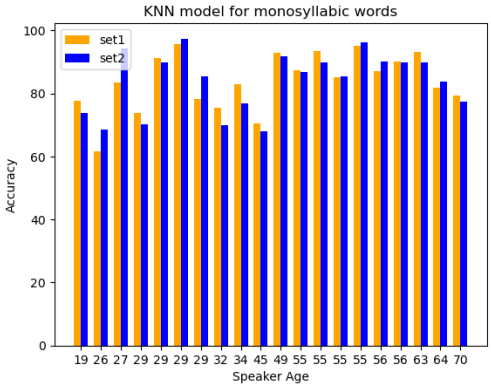


(f) Result of LSTM model on all word pairs

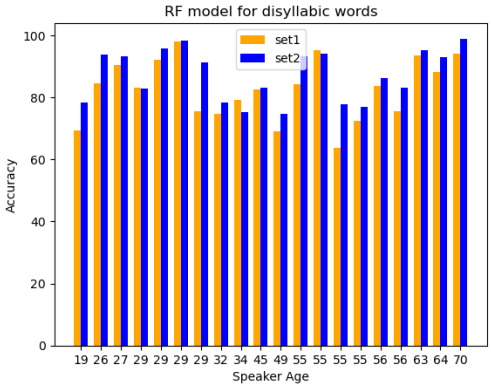
Fig. 6.8: Tone recognition results using SVM and LSTM classifier utilizing feature Set1 and Set2 on different data types for each speaker.



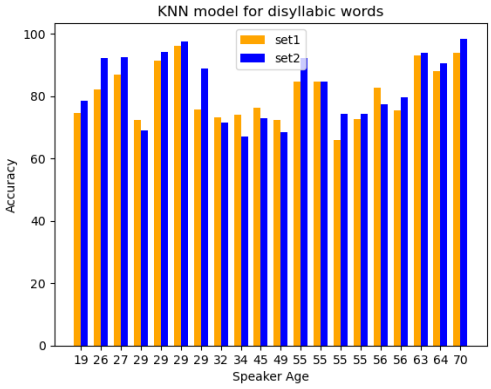
(a) Result of RF model on monosyllabic words



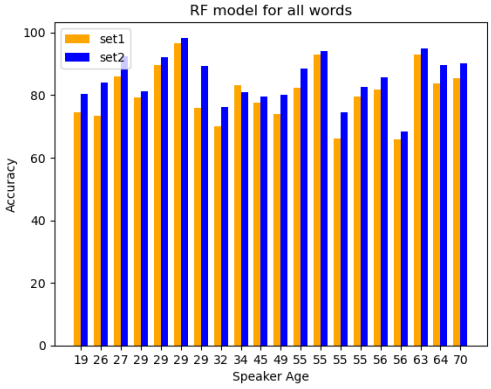
(b) Result of KNN model on monosyllable



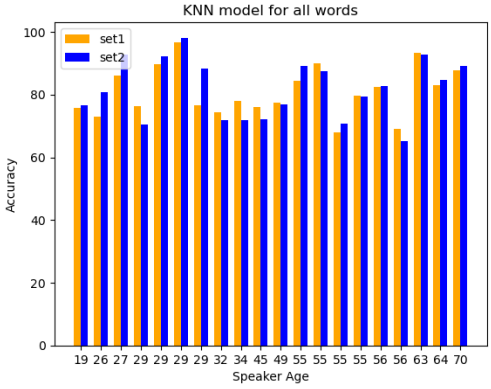
(c) Result of RF model on disyllabic words



(d) Result of KNN model on disyllabic words



(e) Result of RF model on all word pairs



(f) Result of KNN model on all word pairs

Fig. 6.9: Tone recognition results using RF and KNN classifier utilizing feature Set1 and Set2 on different data types for each speaker.

age 70, Fig. 6.8c) for disyllables and 98.46% (Speaker ID 8, age 29, Fig. 6.8e) for all pairs using SVM.

Table 6.9: Tone classification results in speaker dependent scenario

Data Type	Feature Set	Average mean accuracy(%) for 20 speakers			
		SVM	LSTM	RF	KNN
Monosyllable	Set1	85.12	84.84	84.23	83.784
	Set2	87.12	87.30	**87.68**	83.781
Disyllable	Set1	82.35	83.25	82.49	80.78
	Set2	86.46	**87.34**	87.14	82.90
All word	Set1	80.82	81.00	80.50	80.85
	Set2	84.68	84.74	**85.14**	81.67

Table 6.9 displays the average tone recognition mean accuracy achieved by the four classifiers in the speaker-dependent approach across various utterance types. Notably, Set2 demonstrates superior performance to Set1, attaining the highest accuracy values of 87.68%, 87.34% and 85.14% for monosyllables, disyllables and all word pairs, respectively. The RF classifier consistently delivers elevated results on average when applied to the feature Set2.

6.4 Conclusion

This study analyzes tone recognition systems for the tonal language Manipuri. The speech corpus developed in this work is a handcrafted database consisting of 17,837 samples of 54 pairs of tonal contrast words from 20 native speakers. Two feature sets are applied to detect Manipuri's falling and level tones. Five features are extracted from the pitch contour in the first feature set and for the second set, along with five pitch-derived features, shimmer and jitter parameters are added to the features vector. Incorporating the feature sets with a low-complexity machine learning classifier, SVM, LSTM, RF and KNN makes it well-suited for novice Manipuri tone recognition. Comparative experiments validate the superior performance of the RF model on the second feature set,

highlighting its advantages in terms of high recognition accuracy, simplicity and robust learning capability. It also indicates that the pitch variation between the gender and the variation induced by co-articulation between syllables influence the tone recognition rate of the Manipuri tone. The primary applications of this approach are envisioned to benefit individuals with disabilities and children learning Manipuri as their second language. Specifically, it can contribute to developing an application designed to aid native Manipuri speakers in learning Manipuri systematically, focusing on accurate pronunciation and improving their language skills.

7

Conclusion and Future Work

This thesis represents a significant step forward in understanding and processing Manipuri tones, contributing valuable insights to the broader field of linguistics and technology. We have achieved several key findings by developing the ManiTo speech corpus, analyzing Manipuri tones and identifying avenues for future research.

First and foremost, creating the ManiTo dataset, comprising over 17,837 samples of isolated Manipuri tonal contrast words collected from 20 native speakers, amounting to a size of 1.03 GB, fills a critical gap in resources for studying Manipuri tones. This development signifies a significant milestone in facilitating research and technology initiatives to understand and process the Manipuri language. Through meticulous recording, annotation and organization, the dataset now serves as a valuable resource for investigating tonal contrasts and patterns within Manipuri speech. Including speakers from diverse age groups and recording environments enhances the dataset's robustness and applicability across various linguistic contexts. Two variants of the dataset have been developed to augment its utility. The first variant, ManiToSpc, comprises spectrograms extracted from the WAV files in the ManiTo corpus, providing a two-dimensional representation of the audio data. The second variant, ManiToScl, consists of scalograms generated from the audio files, offering additional insights into the tonal characteristics of Manipuri speech. These representations pave the way for applying deep learning techniques treating audio data as images, thus fostering future research and analysis opportunities in audio processing and classification.

Our method of analyzing Manipuri tones through amplitude normalization and pitch length normalization has yielded valuable insights into the distinctive features of falling and level tones. Examining pitch contours has enabled us to identify crucial parameters that effectively distinguish between these tonal qualities. Specific characteristics, including the slope and height of the pitch contour, emerge as significant factors in discerning Manipuri tones. Additionally, our exploration of gender-specific speech data underscores the critical role of gender speech normalization in tonal contrast analysis. The findings emphasize the necessity to account for gender-related variations in pitch characteristics, reinforcing the importance of accurate and meaningful comparisons in Manipuri tonal contrast analysis.

Our study on tone detection systems employing acoustic features has revealed the efficacy of integrating shimmer and jitter parameters alongside pitch derived features. Our methodology extracted five features from the pitch contour to form the first feature set. Furthermore, we augmented five pitch-derived features with shimmer and jitter parameters for the second set, culminating in a comprehensive feature vector.

Integration of these feature sets with low-complexity machine learning classifiers such as SVM, LSTM, RF and KNN proved to be well-suited for a Manipuri tone recognition system. Comparative experiments validate the superior performance of the RF model on the second feature set, highlighting its advantages in terms of high recognition accuracy, simplicity and robust learning capability. Furthermore, our analysis indicates that pitch variation between genders and variations induced by co-articulation between syllables influence the tone recognition rate of the Manipuri tones. The primary applications of this approach are envisioned to benefit individuals with disabilities and children learning Manipuri as their second language. Specifically, it can contribute to the development of applications designed to aid native Manipuri speakers in systematic learning, focusing on accurate pronunciation and language skill improvement.

Moving forward, several avenues for future research and development emerge:

- *Expansion and refinement of the ManiTo dataset*: Continuous updates and expansions to the dataset will enhance its relevance and effectiveness in supporting future research and technological advancements in Manipuri language processing.

- *Gender speech normalization*: Further analysis is needed to develop robust methods for gender speech normalization, ensuring accurate and meaningful comparisons between male and female speech data in Manipuri tonal contrast analysis.

- *Exploration of deep learning techniques*: The adoption of deep learning techniques for tone recognition in Manipuri holds promise for advancing the accuracy and efficiency of Manipuri ASR systems. Future research could explore the application of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to leverage the spectrogram and scalogram representations provided by the ManiTo dataset.

- *Development of educational applications*: The development of educational applications designed to aid native Manipuri speakers in learning the language systematically could benefit from the insights gained in this study. Such applications could focus on accurate pronunciation and language skill improvement, catering to individuals with disabilities and children learning Manipuri as their second language.

In conclusion, this thesis marks a significant contribution to the understanding and processing of Manipuri tones. By leveraging technology and linguistic insights, we aim to promote the appreciation and understanding of Manipuri tones while fostering inclusivity and accessibility in language learning and communication.

References

[1] 2017. *Meetei Mayek Tamnaba Lairik*. Textbook, Global Publications.

[2] T. Adegbola A.A. Sosimi and O.A. Fakinlede. 2019. Standard Yorùbá context dependent tone identification using Multi-Class Support Vector Machine (MSVM) . *Journal of Applied Sciences and Environmental Management* 23, 5 (2019). https://doi.org/10.4314/jasem.v23i5.20

[3] Reetz H. Arora V, Lahiri A. 2018 Jan. Phonological feature-based speech recognition system for pronunciation training in non-native language learning.. In *J Acoust Soc Am. 143(1):98. doi: 10.1121/1.5017834. PMID: 29390749*.

[4] Utpal Bhattacharje. 2013. Recognition of the Tonal Words of BODO Language. *International Journal of Recent Technology and Engineering* 1, 6 (January 2013). https://www.ijrte.org/wp-content/uploads/papers/v1i6/F0446021613.pdf

[5] Paul Boersma and Vincent Van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot Int* 5 (01 2001), 341–347.

[6] Paul Boersma and David Weenink. 1992–2022. *Praat: doing phonetics by computer [Computer program] Version 6.1.51*. Retrieved August 20, 2021 from https://www.fon.hum.uva.nl/praat/

[7] Encyclopedia Britannica. 2014. *Phonology*. https://www.britannica.com/science/phonology

[8] Encyclopedia Britannica. 2015. *Manipuri language*. Retrieved April 20, 2023 from https://www.britannica.com/topic/Manipuri-language

[9] Jason Brownlee. 2018. *A Gentle Introduction to k-fold Cross-Validation*. Retrieved January 30, 2023 from https://machinelearningmastery.com/k-fold-cross-validation/

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4960–4964. https://doi.org/10.1109/ICASSP.2016.7472621

[11] Hao Chao, Cheng Song, Bao-Yun Lu, and Yong-Li Liu. 2019. Feature extraction based on DBN-SVM for tone recognition. *Journal of Information Processing Systems* 15 (01 2019), 91–99. https://doi.org/10.3745/JIPS.04.0101

[12] Shobhana L. Chelliah. 1990. Level Ordered Morphology and Phonology in Manipuri. In *Linguistics of the Tibeto-Burman Area*.

[13] Shobhana Lakshmi Chelliah. 1997. *A Grammar of Meithei*. De Gruyter Mouton, Berlin, New York. https://doi.org/doi:10.1515/9783110801118

[14] Suphattharachai Chomphan. 2011. Analytical Study on Fundamental Frequency Contours of Thai Tones using Tone-Geometrical Model. *Journal of Computer Science* 7 (03 2011), 379–386. https://doi.org/10.3844/jcssp.2011.379.386

[15] Wikipedia contributors. [n. d.]. *Meitei Language*. https://en.wikipedia.org/wiki/Meitei_language [Online; accessed 5 February, 2024; last updated 16 January, 2024].

[16] Hajarimayum Subadani Devi. 2004. Loanwords in Manipuri and their impact. In *Linguistics of the Tibeto-Burman Area, Volume 27.1 Spring*.

[17] Abhishek Dey, Wendy Lalhminghlui, Priyankoo Sarmah, K. Samudravijaya, S. R. Mahadeva Prasarma, Rohit Sinha, and S. R. Nirrnala. 2017. Mizo Phone Recognition System. In *2017 14th IEEE India Council International Conference (INDICON)*. 1–5. https://doi.org/10.1109/INDICON.2017.8487726

[18] H. Dolen. 2016. *Manipuri Dictionary A Comprehensive Manipuri to English Learner's Dictionary in Meitei Mayek and Bengali Scripts* (3 ed.). Hodamba Publications. ISBN - 970-81-922950-2-2.

[19] Sushanta Kabir Dutta, Salam Nandakishor, and L Joyprakash Singh. 2017. A comparative study on feature dependency of the Manipuri language based phonetic engine. In *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*. 5–10. https://doi.org/10.1109/CSCITA.2017.8066533

[20] Mireia Farrús and Javier Hernando. 2009. Using Jitter and Shimmer in speaker verification. *Signal Processing, IET* 3 (08 2009), 247 – 257. https://doi.org/10.1049/iet-spr.2008.0147

[21] Mireia Farrús, Javier Hernando, and Pascual Ejarque. 2007. Jitter and shimmer measurements for speaker recognition. *Proceedings of the Interspeech 2007*, 778–781. https://doi.org/10.21437/Interspeech.2007-147

[22] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. 2014. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering* 2, 1 (2014), 602–609. https://doi.org/10.1080/21642583.2014.956265 arXiv:https://doi.org/10.1080/21642583.2014.956265

[23] Qian-Jie Fu, Fan-Gang Zeng, Robert V. Shannon, and Sigfrid D Soli. 1998. Importance of tonal envelope cues in Chinese speech recognition. In *The Journal of the Acoustical Society of America*. 505–510. https://doi.org/10.1121/1.423251

[24] Hiroya Fujisaki. 1983. *Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing*. Springer New York, New York, NY, 39–55. https://doi.org/10.1007/978-1-4613-8202-7_3

[25] Qiang Gao, Shutao Sun, and Yaping Yang. 2019. ToneNet: A CNN Model of Tone Classification of Mandarin Chinese. In *Interspeech*. https://api.semanticscholar.org/CorpusID:202727233

[26] Bhupesh Gha. 2021. Automatic Language Identification of Spoken Hindi and Manipur. *International Journal of Innovative Research in Engineering and Management (IJIREM)* 8 (November 2021). https://ijirem.org/DOC/109-automatic-language-identification-of-spoken-hindi-and-manipuri.pdf

[27] Parismita Gogoi, Abhishek Dey, Wendy Lalhminghlui, Priyankoo Sarmah, and S R Mahadeva Prasanna. 2020. Lexical Tone Recognition in Mizo using Acoustic-Prosodic Features. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6458–6461. https://aclanthology.org/2020.lrec-1.795

[28] Parismita Gogoi, Sishir Kalita, Wendy Lalhminghlui, Priyankoo Sarmah, and S. R. M. Prasanna. 2021. Learning Mizo Tones from F0 Contours Using 1D-CNN. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings* (St Petersburg, Russia). Springer-Verlag, Berlin, Heidelberg, 214–225. https://doi.org/10.1007/978-3-030-87802-3_20

[29] Jyoti Guglani and A.N. Mishra. 2020. Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit. *Applied Acoustics* 167 (2020), 107386. https://doi.org/10.1016/j.apacoust.2020.107386

[30] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. KNN Model-Based Approach in Classification. (2003), 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

[31] Aditya Bihar Kandali Gurumayum Robert Michael. 2022. Emotion Recognition from Manipuri Language Using MFCC and Convolution Neural Network. *International Journal For Research in Applied Science and Engineering Technology,(IJRASET)* (2022). https://doi.org/10.22214/ijraset.2022.47463

[32] Loitongbam Gyanendro Singh, Lenin Laitonjam, and Sanasam Ranbir Singh. 2016. Automatic Syllabification for Manipuri language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Yuji Matsumoto and Rashmi Prasad (Eds.). The COLING 2016 Organizing Committee, Osaka, Japan, 349–357. https://aclanthology.org/C16-1034

[33] Chen Hangting, Pengyuan Zhang, Haichuan Bai, Qingsheng Yuan, Xiuguo Bao, and Yonghong Yan. 2018. Deep Convolutional Neural Network with Scalogram for Audio Scene Modeling. 3304–3308. https://doi.org/10.21437/Interspeech.2018-1524

[34] Pauthang Haokip. 2011. The languages of manipur: A case study of the Kuki-Chin languages. *Linguistics of the Tibeto-Burman Area* 34 (April 2011), 85–118.

[35] Marti A. Hearst, Susan T. Dumais, Edger Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications* 13, 4 (1998), 18–28. https://doi.org/10.1109/5254.708428

[36] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[37] Hao Huang, Ying Hu, and Haihua Xu. 2017. Mandarin tone modeling using recurrent neural networks. *ArXiv* abs/1711.01946 (2017). https://api.semanticscholar.org/CorpusID:32728103

[38] John D. Hunter. 2007. *Matplotlib: A 2D graphics environment*. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.specgram.html.

[39] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[40] Eric Jones, Travis Oliphant, and Pearu Peterson. 2001. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org

[41] Arshpreet Kaur and Amitoj Singh. 2016. Power-Normalized Cepstral Coefficients (PNCC) for Punjabi Automatic Speech Recognition using phone based modelling in HTK. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. 372–375. https://doi.org/10.1109/ICATCCT.2016.7912026

[42] Natthawut Kertkeidkachorn, Proadpran Punyabukkana, and Atiwong Suchato. 2015. Acoustic Features for Hidden Conditional Random Fields–Based Thai Tone Classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 15, 2, Article 9 (dec 2015), 26 pages. https://doi.org/10.1145/2833088

[43] Bipul Syam Purkayastha Kh Raju Singha and Kh Dhiren Singha. 2012. Part of Speech Tagging in Manipuri with Hidden Markov Model. *International Journal of Computer Science (IJCSI)* 9, 2 (November 2012). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=61269edc218657f89014ee616324773de9f689ca

[44] A. G. Khan. 1987. *A Contrastive Study of Manipuri (Meiteilon) and English Phonology*. Ph. D. Dissertation. Guwahati University.

[45] Nepoleon Keisham Biakchungnunga Varte Sivaji Bandopadhyay Kishorjit Nongmeikapam, Chiranjiv Chingangbam. 2014. Chunking In Manipuri Using CRF.

In *International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3* . https://airccse.org/journal/ijnlc/papers/3314ijnlc12.pdf

[46] Mahabir L. 1982. A contribution to the study of tone in Manipuri. In *Master's thesis, Deccan College Postgraduate and Research Institute, Pune*.

[47] Jingjiao Li, Xiaodong Xia, and Shusheng Gu. 1999. Mandarin four-tone recognition with the fuzzy C-means algorithm. In *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315)*, Vol. 2. 1059–1062 vol.2. https://doi.org/10.1109/FUZZY.1999.793100

[48] Tze Fen Li. 2003. Speech recognition of mandarin monosyllables. *Pattern Recognition* 36, 11 (2003), 2713–2721. https://doi.org/10.1016/S0031-3203(03)00135-3

[49] Lih-Cherng Liu, Wu-Ji Yang, Hsiao-Chuan Wang, and Yueh-Chin Chang. 1989. Tone recognition of polysyllabic words in Mandarin speech. *Computer Speech and Language* 3, 3 (1989), 253–264. https://doi.org/10.1016/0885-2308(89)90021-1

[50] Mishaim Malik, Muhammad Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications* 80 (03 2021), 1–47. https://doi.org/10.1007/s11042-020-10073-7

[51] Maleerat Maliyaem, Supot Nitsuwat, and Choochart Haruechaiyasak. 2010. Improving ASR for Continuous Thai Words Using ANN/HMM. 247–256.

[52] Minesh Mathew, Ajeet Kumar Singh, and CV Jawahar. 2016. Multilingual OCR for Indic scripts. In *12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, Santorini, Greece, 186–191.

[53] J. Muradeli. 2020. ssqueezepy. https://github.com/OverLordGoldDragon/ssqueezepy/. [Online].

[54] Kshetrimayum N. 2010. A comparative study of Meetei Mayek: from the inscribed letterform to the digital typeface. *Unpublished Masters Dissertation. University of Reading. Reading, UK* (2010).

[55] Hong Quang Nguyen, Pascal Nocera, Eric Castelli, and Van Loan Trinh. 2008. A novel approach in continuous speech recognition for Vietnamese, an isolating tonal language. In *Proc. Interspeech 2008*. 1149–1152. https://doi.org/10.21437/Interspeech.2008-349

[56] Quoc Bao Nguyen, Tat Thang Vu, and Chi Mai Luong. 2016. The Effect of Tone Modeling in Vietnamese LVCSR System. *Procedia Computer Science* 81 (2016), 174–181. https://doi.org/10.1016/j.procs.2016.04.046 SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

[57] Quang H. Nguyen and Tuan-Dung Cao. 2020. A novel method for recognizing Vietnamese voice commands on smartphones with support vector machine and convolutional neural networks. *Wireless Communications and Mobile Computing* 9 (2020). https://doi.org/10.1155/2020/2312908

[58] Van Huy Nguyen, Chi Mai Luong, and Tat Thang Vu. 2015. Tonal phoneme based model for Vietnamese LVCSR. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. 118–122. https://doi.org/10.1109/ICSDA.2015.7357876

[59] Vu Tat Thang Do Quoc Truong Nguyen Van Huy, Luong Chi Mai. 2014. Vietnamese recognition using tonal phoneme based on multi space distribution. *Journal of Computer Science and Cybernetics* 30, 1 (2014), 28–38. https://doi.org/10.15625/1813-9663/30/1/3553

[60] Jane Oruh, Serestina Viriri, and Adekanmi Adegun. 2022. Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition. *IEEE Access* 10 (2022), 30069–30079. https://doi.org/10.1109/ACCESS.2022.3159339

[61] Tanvina Patel, Krishna DN, Noor Fathima, Nisar Shah, Mahima C, Deepak Kumar, and Anuroop Iyengar. 2018. An Automatic Speech Transcription System for Manipuri Language. In *Proc. Interspeech 2018*. 2388–2389. https://www.isca-archive.org/interspeech_2018/patel18b_interspeech.pdf

[62] Tanvina Patel, Krishna DN, Noor Fathima, Nisar Shah, Mahima C, Deepak Kumar, and Anuroop Iyengar. 2018. Development of Large Vocabulary Speech Recognition System with Keyword Search for Manipuri. In *Proc. Interspeech 2018*. 1031–1035. https://doi.org/10.21437/Interspeech.2018-2133

[63] Linkai Peng, Wang Dai, Dengfeng Ke, and Jinsong Zhang. 2021. Multi-Scale Model for Mandarin Tone Recognition. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362063

[64] Siripong Potisuk, Mary P. Harper, and Jack Gandour. 1999. Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method. *IEEE Transactions on Speech and Audio Processing* 7, 1 (1999), 95–102. https://doi.org/10.1109/89.736336

[65] T. Lakshmi Priya, N.R. Raajan, N. Raju, P. Preethi, and S. Mathini. 2012. Speech and Non-Speech Identification and Classification using KNN Algorithm. *Procedia Engineering, INTERNATIONAL CONFERENCE ON MODELLING OPTIMIZATION AND COMPUTING* 38 (2012), 952–958. https://doi.org/10.1016/j.proeng.2012.06.120

[66] Prof. Bipul Syam Purkayastha Sagolsem Poireiton Meitei, Shantikumar Ningombam. 2012. An Analysis Towards the Development of Electronic Bilingual Dictionary (Manipuri-English) -A Report. In *International Journal of Computer Science and Information Technologies*, Vol. 3. https://ijcsit.com/docs/Volume%203/Vol3Issue2/ijcsit2012030226.pdf

[67] Loitongbam Sanayai Meetei, Laishram Rahul, Alok Singh, Salam Michael Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021. An Experiment on Speech-to-Text Translation Systems for Manipuri to English on Low Resource Setting. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*. NLP Association of India (NLPAI), National Institute of Technology Silchar, Silchar, India, 54–63. https://aclanthology.org/2021.icon-main.8

[68] Biswajit Dev Sarma, Priyankoo Sarmah, Wendy Lalhminghlui, and S. R. Mahadeva Prasanna. 2015. Detection of Mizo tones. In *Proc. Interspeech 2015*. 934–937. https://doi.org/10.21437/Interspeech.2015-15

[69] Scikit-learn. 2011. *Cross-validation: evaluating estimator performance*. Retrieved April 30, 2023 from https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-iterators

[70] Usha Shastri and Ajith Uppunda. 2014. Production and perception of lexical tones in Manipuri language. *Journal of Advanced Linguistic Studies* 1-2 (12 2014), 216–231. https://www.researchgate.net/publication/269697543_Production_and_perception_of_lexical_tones_in_Manipuri_language

[71] Jonathon Shlens. 2014. *A Tutorial on Principal Component Analysis*. Google Research. Mountain View, CA 94043. https://arxiv.org/pdf/1404.1100.pdf

[72] G SijiGeorgeC and B.Sumathi. 2020. Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction. *International Journal of Advanced Computer Science and Applications* 11 (2020). https://api.semanticscholar.org/CorpusID:222433720

[73] Amitoj Singh and Virender Kadyan. 2020. Automatic Speech Recognition System for Tonal Languages: State-of-the-Art Survey. *Archives of Computational Methods in Engineering* 28 (February 2020). https://doi.org/10.1007/s11831-020-09414-4

[74] Chungkham Yashwanta Singh. 2019. *Manipuri Grammar* (2nd. ed.). Rajesh Publications.

[75] H. Dolen Singh and S. Imoba Singh. 2007. Exotic Phonemes: A Study of Manipuri Phonemes. *Language in India* 7 (August 2007). http://www.languageinindia.com/aug2007/manipuriexotic.pdf

[76] Leihaorambam Singh, Kabita Thaoroijam, and Pradip Das. 2007. Written Manipuri (Meiteiron): From Phoneme to Grapheme. *Language in India* 7 (June 2007). http://www.languageinindia.com/june2007/meiteiphoneme.pdf

[77] Fengxi Song, Zhongwei Guo, and Dayong Mei. 2010. Feature Selection Using Principal Component Analysis. In *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, Vol. 1. 27–30. https://doi.org/10.1109/ICSEM.2010.14

[78] Krittakom Srijiranon and Narissara Eiamkanitchat. 2015. Thai speech recognition using Neuro-fuzzy system. In *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. 1–6. https://doi.org/10.1109/ECTICon.2015.7207075

[79] Syntrillium Software Corporation. 2000. *Cool Edit 2000*. https://archive.org/details/cooledit2000v1

[80] Meiraba Takhellambam. 2014. Tones in Meiteilol: A Phonetic Description. In *Language in India*. http://www.languageinindia.com/july2014/meirabameiteiloltones1.pdf

[81] Adrian Tam. 2021. *Principal Component Analysis for Visualization*. Retrieved April 30, 2023 from https://machinelearningmastery.com/principal-component-analysis-for-visualization/

[82] Lin Tang and Junxun Yin. 2006. Mandarin Tone Recognition Based on Pre-Classification. In *2006 6th World Congress on Intelligent Control and Automation*, Vol. 2. 9468–9472. https://doi.org/10.1109/WCICA.2006.1713835

[83] João Teixeira, Carla Oliveira, and Carla Lopes. 2013. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology* 9 (12 2013), 1112–1122. https://doi.org/10.1016/j.protcy.2013.12.124

[84] Nipon Theera-Umpon, Suppakarn Chansareewittaya, and Sansanee Auephanwiriyakul. 2011. Phoneme and tonal accent recognition for Thai speech. *Expert Systems with Applications* 38, 10 (2011), 13254 – 13259. https://doi.org/10.1016/j.eswa.2011.04.142

[85] Purna Chandra Thoudam. 1980. A Grammatical Sketch of Meiteiron. In *Thesis of Doctor of Philosophy, Jawaharlal Nehru University, New Delhi*.

[86] Haydée F. Wertzner, Solange Schreiber, and Luciana Amaro. 2005. Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. *Brazilian Journal of Otorhinolaryngology* 71, 5 (2005), 582–588. https://doi.org/10.1016/S1808-8694(15)31261-1

[87] Richard Wiese. 1986. The Role of Phonology in Speech Processing. In *Proceedings of the 11th Conference on Computational Linguistics* (Bonn, Germany) *(COLING '86)*. Association for Computational Linguistics, USA, 608–611. https://doi.org/10.3115/991365.991543

[88] Wikipedia. [n. d.]. *Phoneme*. https://en.wikipedia.org/wiki/Phoneme [Online;last updated 8 January, 2024].

[89] Jiang Wu, Stephen A. Zahorian, and Hongbing Hu. 2013. Tone recognition for continuous accented Mandarin Chinese. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7180–7183. https://doi.org/10.1109/ICASSP.2013.6639056

[90] Lonce Wyse. 2017. Audio Spectrogram Representations for Processing with Convolutional Neural Networks. (06 2017), 37–41. https://www.researchgate.net/publication/318030481_Audio_Spectrogram_Representations_for_Processing_with_Convolutional_Neural_Networks

[91] Zhaoqiang Xie and Zhenjiang Miao. 2010. Tone recognition of isolated Mandarin syllables. In *Proceedings of the 4th International Conference on Image and Signal Processing* (Trois-Rivières, QC, Canada) *(ICISP'10)*. Springer-Verlag, Berlin, Heidelberg, 412–418.

[92] Jiameng Yan, Lan Tian, Xiaoyu Wang, Junhui Liu, and Meng Li. 2023. A Mandarin Tone Recognition Algorithm Based on Random Forest and Features Fusion. In *Proceedings of the 7th International Conference on Control Engineering and Artificial Intelligence* (, Sanya, China,) *(CCEAI '23)*. Association for Computing Machinery, New York, NY, USA, 168–172. https://doi.org/10.1145/3580219.3580249

[93] Moira Yip. 2002. *Tone*. Cambridge Textbooks in Linguistics, Cambridge University Press.

[94] Yaxin Zhang, Anton Medievski, James Lawrence, and Jianming Song. 2002. A study on tone statistics in Chinese names. *Speech Communication* 36, 3 (2002), 267–275. https://doi.org/10.1016/S0167-6393(01)00007-3

[95] Ọdétúnjí Àjàdí ỌDéloBí. 2008. *Recognition of Tones in YorÙbÁ Speech: Experiments With Artificial Neural Networks*. Springer Berlin Heidelberg, Berlin, Heidelberg, 23–47. https://doi.org/10.1007/978-3-540-75398-8_2

Publications

Journal

1. **Thiyam Susma Devi** and Pradip K. Das. 2024."**Disambiguation of Isolated Manipuri Tonal Contrast Word Pairs Using Acoustic Features.**" ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23, 3, Article 44 (March 2024), 18 pages. https://doi.org/10.1145/3643830

Conferences

1. **Thiyam Susma Devi** and Pradip K. Das. "**Speech Dataset Development for a Low-Resource Tibeto-Burman Tonal Language.**" 2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Delhi, India, 2023, pp. 1-6, doi: 10.1109/O-COCOSDA60357.2023.10482932. https://ieeexplore.ieee.org/document/10482932

2. **Thiyam Susma Devi** and Pradip K. Das. 2022. "**Towards Manipuri Tonal Contrast Disambiguation Using Acoustic Features**" In 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST) pp 1–5. https://doi.org/10.1109/AIST55798.2022.10065089

3. **Thiyam Susma Devi** and Pradip K. Das. 2021. "**Analysis of Manipuri Tones in ManiTo: A Tonal Contrast Database.**" In Proceedings of the 18th International Conference on Natural Language Processing (ICON). NLP Association of India (NLPAI), National Institute of Technology Silchar, Silchar, India, pp 601–605. https://aclanthology.org/2021.icon-main.73

4. **Thiyam Susma Devi** and Pradip K Das. 2021. " **Development of ManiTo: A Manipuri Tonal Contrast Dataset.**" In 2021 3rd International Conference on Artificial Intelligence and Speech Technology, Published in Artificial Intelligence and Speech Technology Amita Dev, S. S. Agrawal and Arun Sharma (Eds.). Springer International Publishing, Cham, pp 255–263, 2022 https://doi.org/10.1007/978-3-030-95711-7_23 (**BEST PAPER AWARD**)