Automatic Dialect Identification in Ao, a Low Resource Language

Moakala Tzudir

Automatic Dialect Identification in Ao, a Low Resource Language

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

MOAKALA TZUDIR

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

JANUARY 2023

Certificate

This is to certify that the thesis entitled "**AUTOMATIC DIALECT IDENTIFICATION IN AO, A LOW RESOURCE LANGUAGE**", submitted by **MOAKALA TZUDIR** (156102018), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by her under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in our opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated: 25-01-2023
Dharwad.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electrical Engineering
Indian Institute of Technology Dharwad
Dharwad-580 011, Karnataka, India.

Dated: 25-01-2023
Guwahati.

Prof. Priyankoo Sarmah
Professor
Dept. of Humanities and Social Sciences
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

My parents

Mrs. Atsungla and Late Er. Jemsung Tzudir

&

My brothers

Dr. Along Tzudir and Mr. Sunep Tzudir

Acknowledgements

This thesis could not have been completed without the support of many individuals and organizations, to whom I am immensely thankful. I would like to begin by expressing my gratitude and appreciation to the Indian Institute of Technology Guwahati for giving me the opportunity to conduct this research and providing financial support until the thesis submission. Also, I want to thank IIT Guwahati for providing state-of-the-art laboratories, computational resources, and other logistical support.

I would like to convey my profound gratitude to Prof. S. R. M. Prasanna and Prof. Priyankoo Sarmah, who supervised my Ph.D. work. I am grateful for their continuous guidance, motivation, and support throughout the journey. The weekly meeting at 8:00 am every Thursday has helped in the progress of my work and the completion of this thesis. I would also like to thank them for their financial assistance in attending several conferences. I owe a debt of gratitude to several people, foremost among them Prof. S. Dandapat, the Chairman of my Doctoral Committee, for encouraging and providing valuable suggestions on my work throughout the years; Prof. Rohit Sinha and Dr. Bidisha Som, member of my Doctoral Committee, for their insightful suggestions and critical feedback that helped to improve my thesis work; Dr. T. Temsunungsang (EFLU Shillong) for his expert feedback on the Ao tones; Dr. Tiasunep Amri (Nagaland University) and Dr. Sentimenla Jamir for translating the Bible passage to Changki and Mongsen dialects.

In the Department of Electronics and Electrical Engineering (EEE), IIT Guwahati, I am grateful to the head of the Department, Prof. Palathinkal Paily Roy for the support and for extending all possible help needed in academic work. In addition, I would like to convey my profound gratitude to all of the EEE's non-teaching personnel, namely, Mr. Mukut Baruah, Mr. Sundeep Borah, Mrs. Krishangi K Bhuyan,

and Mr. Dasarath Das for their unwavering assistance and quick responsiveness to my technical and administrative needs.

My gratitude to all the seniors, Dr. Biswajit Dev Sarma, Dr. Nagaraj Adiga, Dr. Rohan Kumar Das, Dr. Bidisha Sharma, Dr. Banriskhem K. Khonglah, Dr. Rajib Sarma, Dr. Himakshi Choudhury, Dr. Subhasis Mondal, Dr. Vikram C. M., Dr. Sishir Kalita, Dr. Akhilesh Dubey, and Dr. Protima Nomo Sudro, who have provided an excellent research environment.

I am equally grateful for colleagues from the Signal Processing Lab., Signal Informatics Lab., and Electro Medical and Speech Technology (EMST) Lab. I sincerely thank Dr. Shikha Baghel and Mrinmoy Bhattacharjee for their help at different stages of this journey. I highly appreciate the discussions and suggestions received from both of them. I will always cherish the time spent with Protima, Shikha, Mrinmoy, Brij, and Anik in the Signal Processing lab. The endless discussions with my weekly meeting partners, Saswati and Pari after every meeting will always be remembered. I am incredibly grateful to Nama for being there when I initially came for my Ph.D. entrance exam. I am fortunate to have good friends on campus whose friendship has been an immense blessing. My heartfelt appreciation to Ato, Viya, Naro, and Nini for always being there for me. I would like to thank Ato, Alex, and Angelus for the good times spent in campus. I also extend my sincere gratitude to the IITG Naga family.

This work would not have been possible without the support and love from the participants of Changki, Khensa, and Mopungchuket villages. I am forever grateful to each individual who willingly came forward to contribute the speech data during the multiple field visits.

For their unending love and blessings, I dedicate this accomplishment to my mother and my brothers, Along and Sunep. I am appreciative of them for allowing me the freedom to pursue my dreams and for always being there for me. I am

thankful for the love I have received from my two wonderful sisters-in-law, as well as for the joy my nieces and nephews have provided to this otherwise monotonous journey. It would not have been possible for me to make it this far without the support and blessings of my family.

Above all, I thank God for all His blessings!

Moakala Tzudir

Abstract

Dialect Identification (DID) is a significant research problem widely explored in major languages like Arabic, Chinese, and Spanish. DID can serve as a frontend for many applications like Automatic Speech Recognition (ASR) that may require special dialect-specific enhancements for improved performance. This thesis proposes an automatic DID system for Ao, an under-resourced language of India. Ao is a Tibeto-Burman language spoken in Nagaland. It is a tonal language with three lexical tones: high, mid, and low. Chungli, Mongsen, and Changki are the three dialects of Ao that differ in their respective tone assignment on lexical words. Four principal contributions are made in this thesis. The first contribution of this thesis is creating a manually collected and annotated novel speech dataset to foster research on the Ao language. The second contribution of the thesis is a detailed acoustic study of the unexplored tone dynamics of the dialects of Ao. Based on the analysis, a tonal feature (F_0) to capture the dialect-specific tone information is proposed. The DID performance improves when the proposed tonal feature is combined with other spectral features. As the third contribution, this thesis explores three excitation source features in the DID task. The source features studied are Residual Mel Frequency Cepstral Coefficient (RMFCC), Integrated Linear Prediction Residual Log Mel Spectrogram (ILPR-LMS), and Linear Prediction (LP)-gammatonegram. A notable performance improvement is observed when the source information is combined with the vocal tract information. The fourth contribution of this thesis is the exploration of prosody-related characteristics of speech signals. The prosodic features are observed to provide significant performance improvements in classifying the dialects of Ao. The thesis work is concluded by combining all the proposed approaches to build an efficient DID system for Ao. Among many hurdles in studying under-resourced languages like Ao, the need for more data is the most prominent. Nevertheless, the contributions of this thesis may bridge some of those gaps and spur future research in this direction.

Contents

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Figures

List of Tables

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Acronyms

AANN	Auto-Associative Neural Network
ANOVA	Analysis of Variance
Bi-GRU	Bidirectional Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short-Term Memory
BN	Batch Normalization
BNF	Bottleneck Features
CCA	Canonical Correlation Analysis
CHMM	Continuous Density Hidden Markov Model
CMC	Changki Mongsen Chungli
CMVN	Cepstral Mean and Variance Normalization
CNN	Convolutional Neural Network
CRF	Conditional Random Forest
CSVM	Clustered Support Vector Machines
DCT	Discrete Cosine Transform
DID	Dialect Identification
DNN	Deep Neural Network
ECAPA	Emphasized Channel Attention, Propagation and Aggregation
EER	Equal Error Rate
EM	Expectation–Maximization
ERF	Extreme Random Forest

F1, F2, F3, F4	First, Second, Third and Fourth Formants
FS	Frame Selection
FSD	Frame Selection Decoding
GCI	Glottal Closure Instants
GMBM	Gaussian Mixture Bigram Model
GMM	Gaussian Mixture Model
H	High
HMM	Hidden Markov Model
HNR	Harmonic-to-Noise Ratio
ILPR	Integrated Linear Prediction Residual
ILPR-LMS	ILPR Log Mel Spectrogram
ITU	International Telecommunication Union
KLD	Kullback-Leibler Divergence
KNN	K-Nearest Neighbours
L	Low
LID	Language Identification
LLD	Low-Level Descriptors
LME	Linear Mixed Effects
LMS	Log Mel Spectrogram
LP	Linear Prediction
LPC	Linear Prediction Coefficients
LPCC	Linear Prediction Cepstral Coefficients
LSTM	Long Short-Term Memory
M	Mid
MANOVA	Multivariate Analysis of Variance
MCE	Minimum Classification Error
MFCC	Mel Frequency Cepstral Coefficients

MLE	Maximum Likelihood Estimation
MS	Mixture Selection
MSA	Modern Standard Arabic
NFC	Neuro Fuzzy Classifier
PRLM	Phone Recognition Language Modeling
RBF	Radial Basis Function
ReLU	Rectified Linear activation Unit
RF	Random Forest
RMFCC	Residual Mel-Frequency Cepstral Coefficients
RMS	Root Mean Square
SDC	Shifted Delta Cepstral
SFF	Single Frequency Filtering
SGMM	Subspace Gaussian Mixture Model
STFT	Short Term Fourier Transform
SVM	Support Vector Machines
TBU	Tone Bearing Units
TCN	Temporal Convolution Neural Network
TDNN	Time-Delay Neural Network
UBM	Universal Background Model
VT	Vocal Tract
VQT	Voice Quality and Temporal
XGB	Extreme Gradient Boosting
ZCR	Zero-Crossing Rate
ZFF	Zero-Frequency Filtering
ZTW	Zero-Time Windowing

Chapter 1

Introduction

1.1 Dialect identification: An overview

In today's technologically driven environment, speech-based technologies are essential to a wide range of applications. In recent decades, there has been a great deal of interest in the automatic extraction of information from speech signals. One of the key areas of study in the field of speech research is Dialect Identification (DID) task [3]. DID is one of the major research topics in the speech research community because of its importance in Automatic Speech Recognition (ASR) tasks. The objective of DID task is to distinguish one dialect from the other within the same language family [3].

In simple terms, a dialect can be termed as a speaker's pronunciation and vocabulary variation based on geographical regions [4]. Dialectal variations may also result in syntactic and morphological differences. Dialectal regions may be distributed over a large area. For example, in case of the Arabic language, the dialects such as Modern Standard Arabic (MSA), Iraqi Arabic, and Levantine Arabic are standard dialects of different countries. These dialects are used in broadcast news with available written scripts. Various varieties of English, such as American English, British English, Indian English, and Australian English, are considered dialects of English. Again, within these dialects, there may be sub-dialects, as in the case of American English, which is known to have several regional dialects, viz., the Northern and the Southern dialect [5]. For instance, Figure 1-1 shows the generic term used for carbonated drinks based on regional American English [6]. Contrary to these examples, in case of Naga languages, dialectal variations can be observed within a small geographic area, sometimes even within the boundary of a village.

- Northwest & Midwest - "pop"; South - "coke"; Northeast & Southwest - "soda".

As a result, *Chambers et al.* [4] reports that among the dialects of a language, there exhibit variations in grammatical, phonotactic, phonological, and prosodic differences.

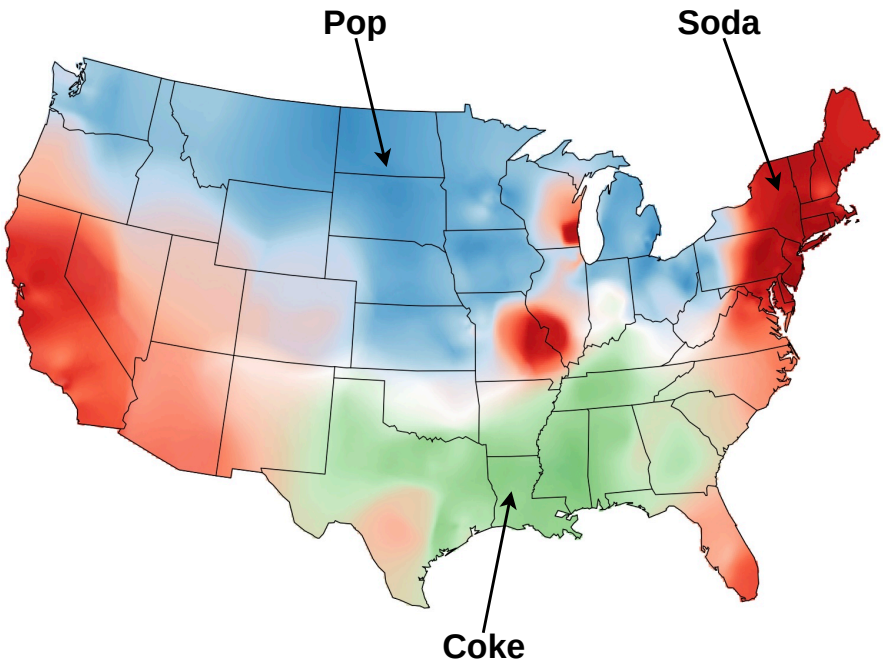


Figure 1-1: Generic term for carbonated drinks in America

According to Wolfram and Schilling [7], dialect refers to "*any variety of a language spoken by a group of people. Also, a language is always expressed through dialects, and to speak a language entails speaking the dialect of that language*" [7]. Dialects, in general, are sub-divisions of a language and the genuine forerunners of language varieties that we have today. While the task of a DID system is to automatically identify one dialect from the other, capturing dialectal variation itself is a formidable challenge [3].

Automatic DID has sparked interest in both academia and industry due to its potential positive impact on society. The automatic DID is expected to enhance human-computer interaction applications [8]. It is also beneficial in providing new services for e-health and telemedicine, particularly for the elderly [8]. In addition, it is helpful for directory assistance or emergency operators through call routing. For example, in applications such as telephone-based assistance systems, when callers do not speak the standard dialect of a particular language, the time it takes to find a good interpreter can be minutes, which can be devastating in an emergency situation.

As a result, a DID that could quickly detect the most likely dialects of incoming speech might reduce the time it takes to find a suitable interpreter by creating a more user-friendly environment for the application's users [9]. As a specific use case, ASR can be modeled with the outcomes of the DID system by enhancing dialect-specific information [10]. This dialect-specific information can be used in smart devices for voice-based interaction applications such as Siri, Alexa, Google Home, etc.

As with the other parts of the world, DID is considered with due importance in the context of the Indian languages. Already DID is integrated into ASR systems of Kannada [11], Telugu [12], and Assamese [13]. However, not much attempt has been made to develop DID systems in under-resourced and minor languages of India. Despite the fact that these languages exhibit distinct dialectal variations. The language of interest in this thesis, Ao, is an under-resourced language that is known to have three distinct dialects, namely, Chungli, Mongsen, and Changki, that differ in terms of phonemic and prosodic properties. Therefore, this thesis is an attempt to devise a methodology to identify the three dialects of Ao automatically. Considering that, the following sections provide a background of the Ao language and the challenges in automatic dialect identification.

1.2 Background of Ao language

For centuries, India has been a homeland to a diverse range of religions, castes, cultures, languages, and dialects, with the world's second-largest population of over 1.21 billion people as per the 2011 Census of India. India is a multilingual country with several languages and dialects, reflecting the country's rich cultural variety and heritage. According to the 2011 Census of India, there are 121 languages and 270 mother tongues in total. Out of which, 22 languages are mentioned in the 8^{th} schedule of the Indian constitution. The geographical area of India is 3.287 million km^2, which

is $\approx 1/3^{rd}$, the geographical area of the United States. The North-East part of India comprises eight states covering $262,179\ km^2$ of India with 3.76% of the total population. There are four major language families in India: Indo-Aryan, Dravidian, Austroasiatic, and Sino-Tibetan (particularly Tibeto-Burman).

Ao is an under-resourced, Tibeto-Burman language spoken in the Northern part of Nagaland (see map in Figure 3.1) in the North-East of India [14]. It covers an area of 1719 km^2 of India. There are three distinct dialects of Ao, viz., Chungli, Mongsen, and Changki [15, 2]. It is a tone language and is reported to have three lexical tones, namely, high (H), mid (M), and low (L) [1, 2]. As per the Census of India 2011, the resident population of Ao in Nagaland is 227,000 [16]. Among the three dialects of Ao, the Chungli dialect is considered the standard variety of the language. A Roman script based on the Chungli dialect was developed for writing Ao by the Christian missionaries in the nineteenth century. As spoken and textual resources in the language are scarce, speech analysis and modeling related work in the Ao dialects is challenging. Moreover, as Chungli is considered the standard variety of the language, all text materials are written based on the Chungli dialect, and hence, the other two dialects are underrepresented in Ao texts. As a result of that, Mongsen and Changki speakers also read and write in the Chungli dialect.

Despite of Ao being an under-resourced language, several researchers have studied the different aspects of the language, considering the fact that it is spoken by one of the major tribes of Nagaland. Accordingly, there are a few works available on the Chungli dialect, including grammatical descriptions by [17, 18] and [19]. There is also an early dictionary, with an updated version in 2013, without the tone markings [20]. The tones in the Chungli dialect are also described in [21]. However, none of these works provide an acoustical analysis of the tones. Apart from Chungli, there are a few works available for the Mongsen dialect [1, 15, 22, 23]. However, the least documented dialect with a very few available works is the Changki dialect [2].

In tone languages, it is imperative that every syllable of the language is associated with one of the lexical tones in the language. While the quality and number of tones in the dialects of a tone language may be the same, their assignment may vary according to dialectal variations [24]. As far as the phonological description of the Ao dialects is concerned, there are some exhaustive descriptions of the Chungli and Mongsen dialects [2, 1, 22]. In case of the Mongsen dialect, [1] and [15] report that this dialect has three lexical tones, High (H), Mid (M), and Low (L). The descriptions provided in these works are based on the authors' impressionistic judgments and some limited acoustic analysis of the tones in the language. *Coupe et al.* [1] discusses the acoustic and perceptual characteristics of tones in the Mongsen dialect, where the differences among tone categories are very small in terms of F_0. However, the perception tests demonstrate that these small differences in F_0 are perceivable to the native speakers leading to successful tone identification. As Ao is predominantly a disyllabic language, minimal sets for the three Ao tones are not found in monosyllables. Hence, the examples in Table 1.1 show tone contrasts in the disyllable /təmaŋ/ [1]. It is noticed from the table that for the word /təmaŋ/, the meaning of the word differs corresponding to the change in tones. Ao being a tonal language, assignment of different tones to the two syllables of the word /təmaŋ/ changes the meaning. For example as shown in the Table when the first syllable, /tə/is assigned a Mid (M) tone and the second syllable /maŋ/ is assigned a High (H) tone the meaning is 'all'. Similarly assignment of various combinations of the three tones in Ao namely, High (H), Mid (M), and Low (L) result in distinct meanings.

It is also confirmed in [2] that all the three dialects of Ao, viz., Chungli, Mongsen, and Changki, have three lexical tones, High (H), Mid (M), and Low (L). However, the assignment of the tones may be different across the dialects of Ao, even for the same lexical item. Figure 1-2 shows the canonical pitch contours of the three Ao tones in the Chungli dialects in the mid 80% of the total duration of the tones, calculated

Table 1.1: Tones in Mongsen Ao [1].

Disyllable	Tones	Gloss
/təmaŋ/	MH	'all'
	MM	'body'
	LM	'believe'
	HL	'dark'

at every 2% of the total duration. Canonical pitch contours are the plots that show the standard tones of a tone language. It is usually plotted using F_0 values from the vowel regions. This Figure is plotted using six minimal sets of disyllabic words, which are discussed in detail in chapter 3.

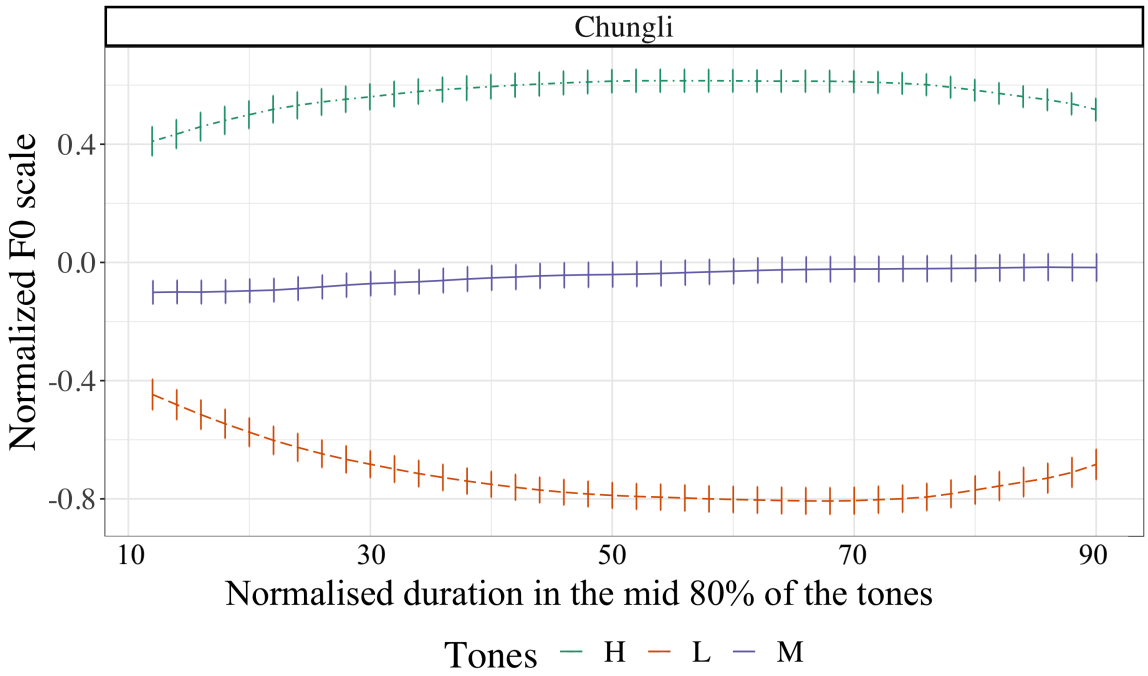


Figure 1-2: Canonical pitch contours of three Ao tones in the Chungli dialect.

For example, the tone assignment in the Changki dialect of Ao is different from the Mongsen and the Chungli dialects for the words in Table 1.2. In the examples provided in Table 1.2, we can see that the Low tone of the Mongsen and Chungli

dialects corresponds to the High tone of the Changki dialect and the High tone of the Mongsen and Chungli dialects corresponds to the Low tone of the Changki dialect as reported in [2]. However, the tone alteration across dialects may not be systematic.

Table 1.2: Tones in Ao dialects [2].

Chungli	Mongsen	Changki	Gloss
azək - HL	alik - HL	alik - LH	'necklace'
akuŋ - HH	akuŋ - MM	akuŋ - HL	'shrimp'
ita - LL	lata - LL	lata - HH	'moon'
təkha - HL	tək^{h}ət - HL	tək^{h}ət - LH	'hand'
miim - HH	məzəm - HH	məzəm - LL	'poison'

1.3 Challenges in DID

In a general sense, a language identification (LID) task and a DID task are based on similar principles. However, building a DID system is considered more difficult than LID system as the boundaries across languages are easier to recognize perceptually and are generally quite distinct. In contrast, dialectal differences are more subtle and may be less salient [25]. Additionally, dialects are closely related to each other as they usually don't differ in their written forms, and they share similar phonemic features. Likewise, overlaps in vocabulary and phonetic features are more common across two distinct dialects of a specific language than across two distinct languages [26]. Considering this, building an effective DID system is viewed as more challenging than building a LID system [27, 10].

Apart from that, the Chungli dialect is known as the standard dialect of the language [15, 2]. As a result, all text materials of the language are written in the Chungli dialect. Therefore, speech modeling and analysis in Ao dialects become very challenging due to the inadequate amount of resources for the other two dialects.

1.4 Motivation of the current work

DID is one of the most emerging research topics in speech [3]. In particular, voice-controlled electronic devices have provided a more comprehensive range of applications by using the outcomes of DID systems. As per the literature, the analysis of the signal processing aspects in the Ao language is not available to the best of our knowledge. Hence, this thesis aims to build an Ao DID system as the Ao language is spoken by one of the most populated tribes in Nagaland, India.

As Ao is a tonal language, the tone dynamics are yet to be known for each dialect. Also, the acoustic study is conducted only in the Mongsen dialect. In addition, the perceptual study is reported only in the Chungli and Changki dialects. As a result, this dissertation investigates the acoustic analysis of tones in the three dialects of Ao. Following that, the tonal information is utilized to capture dialectal variations across the three dialects in Ao DID system.

Next, considering the importance of tonal information in Ao, we presume that excitation source characteristics will play a vital role in determining dialectal distinctions in the three dialects of Ao. It is reported that the aspects of speech production and excitation are different across each sound unit [28]. Also, there are differences in tone assignment across the three dialects. These variations may aid in classifying the three dialects in Ao DID tasks using the excitation source information.

Subsequently, Ao being a tonal language, the differences in the suprasegmental prosodic information due to the contrasts in tones may represent potential distinctions in the three dialects of Ao. Therefore, an automatic Ao DID is designed by exploring the prosodic features. Hence, this thesis is a humble contribution to the community to augment the signal processing research in Ao DID along the lines of widely studied languages like Chinese, Arabic and Spanish.

1.5 Organisation of the thesis

This dissertation aims to automatically identify the three dialects of Ao. Figure 1-3 provides an illustration of the entire Ao DID system developed in this thesis. Initially, a series of pre-processing steps are applied to the speech signal. Following that, the pre-processed speech signal is fed in parallel to the DID systems based on tonal features, excitation source features, and prosodic features. These three systems are then combined to form a final system for classifying the three Ao dialects. The major contributions of this thesis are as follows.

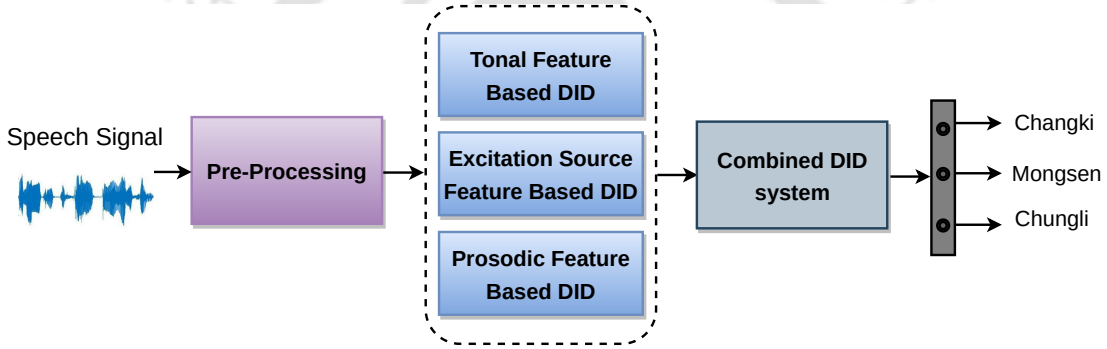


Figure 1-3: Overall Ao DID framework

- Chapter 2 provides a review of the various approaches in DID tasks. The literature is classified into DID tasks in non-tonal languages and DID tasks in tonal languages. This chapter also discusses the scope and motivation of this thesis.

- The speech database, which was solely collected on-site for this research, is described in detail in chapter 3. The Changki Mongsen Chungli-Ao (CMC-Ao) speech corpus contributes three types of Ao speech databases, namely, disyllabic words (DiW), trisyllabic words (TriW), and passage level (PasL). These three datasets were employed to build an automatic dialect identification system in the three dialects of Ao in the following chapters.

- Chapter 4 reports the acoustical analysis of tones in the three dialects of Ao using DiW corpus. Further, TriW corpus is used to see the effectiveness of tonal features (F_0) along with its Δ and $\Delta\Delta$ derivatives to distinguish the three Ao dialects.

- In the Ao DID task, the significance of the excitation source feature, namely, Residual Mel Frequency Cepstral Coefficient (RMFCC) using TriW corpus, is discussed in chapter 5. Additionally, the PasL corpus is used to investigate excitation source parameters such as Integrated Linear Prediction Residual Log Mel Spectrogram (ILPR-LMS), and Linear Prediction (LP)-gammatonegram.

- Chapter 6 discusses the relevance of prosodic information using openSMILE toolkit in classifying the three dialects of Ao. Initially, statistical prosodic features are used to classify Ao dialects in the TriW corpus. Following that, the PasL corpus is used to study the prosodic features of Low Level Descriptors (LLD).

- In chapter 7, a combined Ao DID system is proposed, utilizing the features proposed in chapter 4, chapter 5, and chapter 6. Furthermore, the PasL corpus is annotated at Breath Group (BrG) to examine the proposed Ao DID system from an application perspective.

- Chapter 8 summarizes the primary contributions of the dissertation and provides possible directions for future work.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 2

Dialect Identification - A review

Overview

This chapter provides a review of various works on Dialect Identification (DID). The languages studied in the literature for DID tasks are divided into two categories: DID in non-tonal languages and DID in tonal languages. In most cases, DID tasks are conducted in high-resource languages regardless of the language category. As a result, a section devoted to DID tasks computed on low-resource languages is also included with the intention of contrasting them with the Ao language. The review is followed by a discussion to design an overall framework of an automatic Ao DID system for this dissertation.

2.1 Introduction

A body of work is dedicated to DID systems built in major world languages, such as Arabic, Mandarin Chinese, English, and Spanish. As these languages are spoken over a large geographical area, they have developed several regional dialects over time. For instance, Arabic has several dialects based on the countries and regions it is spoken in, such as Modern Standard Arabic (MSA), Iraqi Arabic, and Levantine Arabic. In this chapter, the languages explored in the literature for DID tasks are classified into two broad categories, namely, DID in non-tonal languages and tonal languages. English is a typical example of a non-tonal language, whereas Mandarin Chinese is a language with tones. In non-tone languages, the pronunciation of a word can change depending on the person's emotion [29]. However, the meaning of the word remains the same. In contrast, in tone languages such as Cantonese Chinese, the same word pronounced differently can carry multiple meanings depending on the change of tones [29]. In Cantonese, there are six different tones. As a result, a word can be uttered in six different tones, consisting of six different meanings [30]. For example, the Cantonese word '/yau/' has the following six meanings [30]:

high level	'worry'
high rising	'paint'
mid level	'thin'
low level	'again'
very low level	'oil'
low rising	'have'

There is also a section dedicated to DID tasks attempted on low-resource languages with the goal of comparing them to the Ao language.

2.2 DID in non-tonal languages

There are numerous works in DID task that deals with non-tonal languages. Several research studies have reported DID systems built for Arabic dialect classification. *Biadsy et al.* reported automatic classification of five Arabic dialects, namely, MSA, Iraqi, Levantine, Gulf, and Egyptian dialects, with an accuracy of 81.6%, using the Phone-Recognition Language Modeling (PRLM) approach [3]. The addition of prosodic features to the model increased the classification accuracy to 86.3% [31]. Several works have reported similar improvements in language and dialect classification after the addition of prosodic features. For example, *Rouas et al.* modeled prosodic features such as, fundamental frequency (F_0), energy, and duration initially by an n-gram language model in the classification of six languages, which yielded an accuracy of 83.5% [32]. They also used prosodic features to classify Arabic spoken in three regions, namely, Occidental (Moroccan, Algerian), Intermediate (Tunisian, Egyptian), and Oriental (Lebanese, Jordanian, Syrian), with an accuracy of 98.0%. Recent studies have reported improved accuracy in DID systems that use Bottleneck Features (BNF) derived from a Deep Neural Network (DNN). These attempts with ASR acoustic modeling setup with a bottleneck layer achieved improved DID performance. *Zhang et al.* [33] showed a reduction of Equal Error Rate (EER) in

Chinese and Arabic DID systems using BNF when compared to a baseline i-vector system only with MFCC features [33]. Later, *Zhang et al.* [34] used Deep Neural Network (DNN) based DID systems on Arabic and Chinese dialects using BNF and autoencoder-based features. The best accuracy was achieved when BNF was fused with the baseline MFCC features. Similar DID attempts using the ADI17 dataset, containing Arabic dialects from 17 countries, showed that MFCC features used on a Convolutional Neural Network (CNN) based system performed with an accuracy of 82% [35]. *Lin et al.* used the same dataset and built a transformer-based DID system with Fbank features where the fusion of CNN and transformer increased the dialect classification accuracy to 86.29% [10].

Considering the dialectal variations in Spanish in continental Europe and South America, several works in DID attempted to classify Spanish dialects. *Zissman et al.* reported automatic classification of Peruvian and Cuban dialects of Spanish from the Miami corpus using the parallel PRLM approach with Mel Frequency Cepstral Coefficient (MFCC) features, resulting with an accuracy of 84% [36]. *Torres-Carrasquillo et al.* used the same dataset with Shifted-Delta Cepstral (SDC) features in the Gaussian Mixture Model (GMM) and reported an EER of 30% [37]. They also used the CallFriend corpus to identify dialects in English, Mandarin, and Spanish, resulting in an EER of 13%. In [38], DID was performed using a hybrid Support Vector Machines-GMM (SVM-GMM) system with MFCC and acoustic-phonetic features to classify three Spanish dialects, namely, Peruvian, Cuban, and Puerto Rican from the Miami corpus. This system showed an overall accuracy of 85.09%. *Huang et al.* also performed DID using the same three Spanish dialects from the Miami corpus and using three English dialects out of eight dialects from the IViE corpus viz., Belfast (Northern Ireland), Cambridge (England), and Cardiff (Wales). They used GMM for Mixture Selection (MS-GMM), Frame Selection (FS-GMM), and Minimum Classification Error (MCE-GMM) [39]. The combination of MCE-FS-GMM gave the best

performance with 82.0% and 93.3% and correct recognition for Spanish and English dialects. Similarly, *Lei et al.* reported GMM with Kullback-Leibler Divergence (KLD-GMM) and Frame Selection Decoding (FSD-GMM) using MFCC features for Arabic, Spanish and Chinese dialects [25]. This system achieved an absolute improvement of 8.7%, 8.6%, and 3.4% in Arabic, Spanish and Chinese dialects when compared to the baseline GMM systems with the Maximum Likelihood Estimation (MLE-GMM) method. The MLE-GMM system achieved an accuracy of 67.7%, 75.8%, and 81.2% in Arabic, Spanish and Chinese dialects, respectively.

DID systems were also explored for American English varieties as spoken in Southern, Western, Northern United States, and New York City, using prosodic and phonotactic features [5]. *Etman et al.* reported that in American English dialect classification, the combined features outperformed the standalone phonotactic feature [5]. *Chittaragi et al.* attempted DID tasks in nine English dialect regions: Belfast, Bradford, Cardiff, Cambridge, Dublin, Leeds, Liverpool, London, and Newcastle [40]. These nine English dialects were classified using MFCC, spectral flux, entropy, pitch, and energy features in SVM, achieving an accuracy of 91.38% for the combined features. Later, they explored the same dataset of nine English dialects to build a DID system using chroma-spectral shape features in SVM, achieving improved performance of 97.52% [41].

Table 2.1: Summary of DID systems related to non-tonal languages

Dialects	**Methodology**	**Accuracy**
5 Arabic dialects: MSA, Iraqi, Levantine, Gulf & Egyptian	PRLM approach	81.6% [3]
	addition of prosodic features in [3]	86.3% [31]
3 Arabic dialect regions: Occidental, Intermediate & Oriental	F_0, energy & duration by an n-gram language model	98.0% [32]
Arabic & Chinese dialects	BNF with i-vector strategy	81.3% & 97.8% for Arabic & Chinese [33]

	BNF in DNN	84.7% & 98.1% for Arabic & Chinese[34]
ADI17 dataset: Arabic dialects from 17 countries	MFCC in CNN	82% [35]
	Fbank features in transformer-based DID system	86.29% [10]
2 Spanish dialects: Peruvian & Cuban	MFCC with parallel PRML approach	84% [36]
	SDC in GMM	EER of 30% [37]
3 Spanish dialects: Peruvian, Cuban & Puerto Rican	MFCC & acoustic-phonetic features in SVM-GMM	85.09% [38]
	MS-GMM, FS-GMM & MCE-GMM	82.0% [39]
Arabic, Spanish & Chinese dialects	MFCC in KLD-GMM & FSD-GMM	absolute improvement of 8.7%, 8.6% & 3.4% in Arabic, Spanish & Chinese [25]
3 English dialect regions: Belfast, Cambridge & Cardiff	MS-GMM, FS-GMM & MCE-GMM	93.3% [39]
9 English dialect regions	MFCC, spectral flux, entropy, pitch & energy in SVM	91.38% [40]
	chroma-spectral shape features in SVM	97.52% [41]
3 English dialects: US, UK & AU from UT-Podcast database	Mel-SFF spectrogram & Mel-STFT spectrogram in SVM	74.90% [42]
	SFF & ZTW with DNNs, CNNs, TCN, TDNN & ECAPA–TDNN	best with SFF for TCN (81.30%), TDNN (81.53%) & ECAPA-TDNN (85.48%) [43]
4 German dialects: Basel, Bern, Lucerne & Zurich	Lexical features in CNN	64.49% F1-score [44]
	character & word n-grams in SVM	66.20% [45]
	character n-grams, word n-grams & word k-skip bigrams in SVM	Best for character n-grams with 62.1% [46]

6 Korean dialects: Seoul/ Gyeong-gi, Gyeong-sang[1], Jeolla, Chung-cheong, Gang-won & Jeju [1]Only Gyeong-sang is a tonal dialect	F_0 & MFCC in attention based Bi-LSTM	68.51% [47]

In [42], DID was performed in three major English dialects: US, UK, and AU from the UT-Podcast database. They used Mel-Single Frequency Filtering (SFF) spectrogram and Mel-Short Term Fourier Transform (STFT) spectrogram in SVM, attaining the dialect classification performance of 74.90%. Subsequently, *Kethireddy et al.* used the same three English dialects for dialect classification in neural networks [43]. They utilized spectrogram, cepstral coefficients, Mel filter-bank energies, and MFCC derived from SFF and Zero-Time Windowing (ZTW) with modern DNN, CNN, Temporal Convolution Neural Networks (TCN), Time-Delay Neural Network (TDNN), and Emphasized Channel Attention, Propagation and Aggregation in TDNN (ECAPA–TDNN). The best performance is achieved with SFF cepstral coefficients for TCN (81.30%), TDNN (81.53%), and ECAPA-TDNN (85.48%).

There are various studies that report DID tasks in German dialects viz., Basel, Bern, Lucerne, and Zurich. Lexical features were used to classify the five German dialects, with CNN acquiring an average F1-score of 64.49% [44]. *Malmasi et al.* also used the five German dialects in DID task, using character n-grams and word n-grams in SVM, achieving an accuracy of 66.20% [45]. Similarly, the German DID system is designed with the five German dialects using character n-grams, word n-grams, and work k-skip bigrams in SVM [46]. The best performance is obtained for character n-grams with an accuracy of 62.1%.

Recent studies in DID task report the classification of six Korean dialects: Seoul/Gyeong-gi, Gyeong-sang, Jeolla, Chung-cheong, Gang-won, and Jeju [47]. Out of these six dialects, Gyeong-sang is the only dialect with tone. *Lee et al.* use F_0 and MFCC features in attention-based Bi-LSTM (Bidirectional Long Short-Term

Memory), which reported 68.51% as the classification performance.

Automatic DID is reported in numerous works for Indian languages. *Rao et al.* describe the identification of five Hindi dialects, namely, Central, Eastern, Western, Northern, and Southern [48]. DID task was attempted with spectral and prosodic features in Auto-Associative Neural Network (AANN) and SVM, achieving an accuracy of 78% and 81% in AANN and SVM, respectively. While in [49], four Hindi dialects viz., Khariboli, Haryanvi, Bhojpuri, and Bagheli were classified using spectral, pitch, and duration features in AANN. The result showed better performance when the duration feature was fused, making duration an important prosodic feature for Hindi dialect classification. The same four dialects were also attempted in [50], using spectral and prosodic features in GMM-SVM, achieving the best accuracy of 88.7%. Apart from Hindi dialects, dialect recognition was conducted in five Kannada dialect regions: Central, Coastal, Hyderabad, Mumbai, and Southern. The five Kannada dialects were classified in [51] using formant frequencies (F1-F3), energy, pitch, and duration in Random Forest (RF), Extreme RF (ERF), and Extreme Gradient Boosting (XGB) algorithms, which reported 76% as the best accuracy in ERF. Subsequently, *Chittaragi et al.* used MFCC, spectral flux, entropy, pitch, and energy in SVM to classify the five Kannada dialects achieving an improved accuracy of 86.25% [40]. The DID system of the five Kannada dialects is further enhanced to 95.60% with chroma-spectral shape features in SVM [41].

2.3 DID in tonal languages

Besides non-tonal languages, there are several attempts at automatic DID in tonal languages. In tone languages, changes in pitch in syllables induce changes in lexical or grammatical meanings [30]. While prosodic features are also used in DID in non-tone languages, these features become more important considering the systematic use of

Table 2.2: Summary of DID systems related to Indian languages

Dialects	**Methodology**	**Accuracy**
5 Hindi dialect regions: Central, Eastern, Western, Northern & Southern	spectral & prosodic features in AANN & SVM	78% & 81% for AANN & SVM [48]
4 Hindi dialects: Khariboli, Haryanvi, Bhojpuri & Bagheli	spectral, pitch & duration in AANN	91% [49]
	spectral, pitch, duration & intensity in GMM-SVM	88.7% [50]
5 Kannada dialect regions: Central, Coastal, Hyderabad, Mumbai & Southern	formants (F1–F3), energy, pitch & duration in RF, ERF & XGB	best in ERF with 76% [51]
	MFCC, spectral flux, entropy, pitch & energy in SVM	86.25% [40]
	chroma-spectral shape features in SVM	95.60% [41]

prosody in tone languages. A majority of works on automatic DID in tone languages come from Chinese languages. These studies have used prosodic features, along with the traditional features, in DID of the tone languages. For instance, *Ma et al.* use pitch flux along with MFCC feature in GMM-based classification to distinguish Mandarin, Cantonese, and Shanghainese dialects [27]. Both Mandarin and Shanghainese have five tones, whereas Cantonese has nine. In this work, it is shown that the error rate in DID is reduced by 30% when pitch flux is combined with the MFCC features. However, it has also been shown that sufficient dialect-specific information may be captured by incorporating only cepstral feature vectors and phonotactic information. In [26], Hidden Markov Model (HMM) based classification is performed to identify three major Chinese dialects: Mandarin, Holo, and Hakka using the cepstral feature vectors and phonotactic information, yielding an accuracy of 89.6%. Nevertheless, in future extensions of their work, prosodic features were also incorporated in the Con-

tinuous Density Hidden Markov Model (CHMM), resulting in a better DID of 93% accuracy [52]. Further, [53] extended the work by using pitch and MFCC features with the Gaussian Mixture Bigram Model (GMBM) based classification, yielding an accuracy of 94.4%. Apart from these Chinese dialects, DID was also performed in North-Chinese, Wu, Guangdong, and Fujian Chinese dialects in *Mingliang et al.* [54], consisting of 20 male and female speakers for each dialect. Clustered SVM (CSVM) with MFCC and SDC features was performed, yielding an accuracy of 92.5%. The work was further extended with the addition of two prosodic features, namely, F_0 and energy in *Mingliang et al.* [55] by using semi-supervised GMM and SVM classification, achieving 92.2% and 93.5% for GMM and SVM, respectively. Most recently, DID task was performed in ten Chinese dialects, namely, Changsha, Hebei, Hefei, Kejia, Minnan, Nanchang, Ningxia, Shan3xi, Shanghai, and Sichuan [56]. The DID system was conducted using 40-dimensional FBank features in CNN-BiGRU (Bidirectional Gated Recurrent Unit) with DNN obtaining an accuracy of 80.13%.

While there are several works on DID in the Chinese dialects, it may be noted that several of these dialects, such as Holo, Hakka, Wu, Guangdongese, and Fujian, are as distinct as different languages, with no mutual intelligibility among them. As a result of this, DID attempts in the Chinese dialects may be more akin to LID attempts. Apart from the Sino-Tibetan languages, there are also attempts at DID in non Sino-Tibetan tone languages such as Punjabi (Indo-European) and Vietnamese (Austroasiatic). In case of Punjabi, *Singh et al.* explore an identification system for three Punjabi dialects, namely, Majha, Malwa, and Doaba using spectral features in RF [57]. Their attempt at Punjabi DID resulted in an accuracy of 98%. Apart from these three Punjabi dialects, *Goyal et al.* included an additional Punjabi dialect: Poadh for Punjabi DID system [58]. They used Subspace Gaussian Mixture Model (SGMM) based classification to identify the four Punjabi dialects using Linear Prediction Cepstral Coefficients (LPCC), F_0 and formants (F1, F2). An accuracy of

Table 2.3: Summary of DID systems related to tonal languages

Dialects	**Methodology**	**Accuracy**
3 Chinese dialects: Mandarin, Cantonese & Shanghainese	pitch flux & MFCC in GMM	error rate reduced by 30% for combined features [27]
3 Chinese dialects: Mandarin, Holo & Hakka	cepstral feature & phonotactic information in HMM	89.6% [26]
	addition of prosodic features in [26] using CHMM	93% [52]
	pitch & MFCC in GMBM	94.4% [53]
4 Chinese dialects: North-Chinese, Wu, Guangdong & Fujian	MFCC & SDC in CSVM	92.5% [54]
	addition of prosodic features (F_0 & energy) in [54] using semi-supervised GMM & SVM	92.2% & 93.5% for GMM & SVM [55]
10 Chinese dialects: Changsha, Hebei, Hefei, Kejia, Minnan, Nanchang, Ningxia, Shan3xi, Shanghai & Sichuan.	40-D FBank features in CNN-BiGRU+DNN	80.13% [56]
3 Punjabi dialects: Majha, Malwa & Doaba	spectral features in RF	98% [57]
4 Punjabi dialects: Majha, Malwa, Doaba & Poadh	LPCC, F_0, formants (F1, F2) in SGMM	87.45% [58]
3 Vietnamese dialects	13D MFCC & F_0 in GMM	70% [59]

87.45% was obtained for the Punjabi DID system. While in case of Vietnamese, *Hung et al.* report a study on Vietnamese DID on embedded systems for three Vietnamese dialects with 13 MFCC coefficients and F_0 features in GMM with an accuracy of 70% [59].

2.4 DID in low-resource languages

Automatic DID attempts are also seen in under-resourced languages. In a non-tonal language such as Assamese, formants (F1-F4) are used to classify four Assamese dialect groups: Eastern, Central, Kamrupi, and Goalpariya, yielding an accuracy of 89.3%, on a Neuro Fuzzy Classifier (NFC)[60]. Automatic DID was also attempted on five dialects of North Sámi, a non-tone language, namely, Inari, Ivalo, Utsjoki, Kautokeino, and Karasjoki [61]. They used acoustic prosodic features in K-Nearest Neighbours (KNN), SVM, RF, Conditional Random Fields (CRF), and LSTM. This study showed that a combination of energy, F_0, and spectral tilt could yield the best accuracy of 60% for CRF in classifying the five dialects of North Sámi [61]. In a recent work, *Devi et al.* reported DID task in Meeteilon, a tonal language [62]. They conducted DID in the three dialects of Meeteilon: Imphal, Kakching, and Sekmai. RF was used to classify the three Meeteilon dialects using formants (F1-F3), F_0, energy, intensity, and segment duration, which yielded an accuracy of 61.57%.

Table 2.4: Summary of DID systems related to low-resource languages

Dialects	**Methodology**	**Accuracy**
4 Assamese dialect groups: Eastern, Central, Kamrupi & Goalpariya	formants (F1-F4) in NFC	89.3% [60]
5 North Sámi dialects: Inari, Ivalo, Utsjoki, Kautokeino & Karasjoki	energy, F_0 & spectral tilt in KNN, SVM, RF, CRF and LSTM	Best: 60% in CRF [61]
3 Meeteilon dialects: Imphal, Kakching & Sekmai	formants (F1-F3), F_0, energy, intensity & segment duration in RF	61.57% [62]

2.5 Discussion and scope of the current work

This chapter discusses the literature explored in DID tasks. The languages researched in DID tasks are classified into two broad parts: non-tonal languages and tonal languages. Irrespective of which category the language falls under, the majority of the DID studies are performed in high-resource languages. For example, MSA, Iraqi Arabic, and Levantine Arabic are Arabic dialects that are extensively used in commercial television with written scripts. As a result, the resources (speech database) are accessible and in the public domain. Moreover, each variety of the dialect is the standard dialect for different countries. However, for a low-resource language like Ao, there are no open-source speech repositories. Also, the written form is only available for the standard variety, i.e., the Chungli dialect. Furthermore, research in the domain of signal processing and speech technology development in the language of Ao has not been explored. This lacuna has formed the basis for the motivation of our work. Hence, this dissertation attempts to bridge the aforementioned gap by developing a DID system in the language.

It is also observed that most of the existing works are based on various aspects of vocal tract information with different classifiers for both non-tone and tone languages. Regardless of the classification system, it is noticed that the system yields better accuracy when the vocal tract information and prosodic features computed from F_0, energy, and duration are combined. Considering tone languages systematically use F_0 to mark lexical tones, it is expected that the addition of pitch features will considerably improve the accuracy of DID in tone languages. As in case of vowels and consonants, tonemes may also have dialectal variations in tone languages. Apart from that, dialect-specific tone assignment patterns may also distinguish one dialect of a tone language from the other [2]. Considering this, chapter 4 is an attempt at DID in Ao language, using tone-specific F_0 features along with its Δ and $\Delta\Delta$ derivatives.

Subsequently, to the best of our understanding, it is noticed that there are no

works in DID tasks that have explored the excitation source information. Given the significance of $F0$ in tone languages, we believe that excitation source information is equally essential for identifying dialectal differences in these languages. According to *Nandi et al.* [28], each sound unit has unique excitation and articulatory source properties. As a result, the impacts of co-articulation and the various ways in which lexical tones are assigned amongst dialects can cause differences in the features of the same sound unit in a tonal language. Hence, in chapter 5, excitation source features such as Residual Mel Frequency Cepstral Coefficient (RMFCC) and Integrated Linear Prediction Residual Log Mel Spectrogram (ILPR-LMS) are used to distinguish the three dialects of Ao [63, 64]. Furthermore, it is stated by *Xu et al.* [65] that both the production process and the auditory perception of tones by humans serve as the foundation for tone identification. The Gammatone filter is thus employed for this work as it is said to be an appropriate model for human perceptual information [66]. Accordingly, the gammatonegram of Linear Prediction (LP) residual is utilized to design an automatic Ao DID system.

Finally, it is noticed in the literature that some works have employed prosodic features such as F_0, energy, duration, and intensity for both non-tonal and tonal languages. However, prosodic features have not been widely researched in the literature. As Ao is a tonal language, suprasegmental properties of the speech signal may capture the change in various components of tones across the three dialects of Ao. Therefore, in chapter 6, the openSMILE toolkit [67] is utilized to extract numerous prosodic features that have not yet been examined in DID tasks in order to evaluate the efficacy of prosodic information in Ao.

Chapter 3

CMC-Ao Corpora

Overview

The three types of speech databases available in the Changki Mongsen Chungli-Ao (CMC-Ao) corpora are described in this chapter. There are three datasets in the CMC-Ao corpora: disyllabic words (DiW), trisyllabic words (TriW), and passage level (PasL). As the resources in the Ao language were not publicly available, the CMC-Ao speech database was created exclusively for this dissertation. In the following chapters, these three datasets were used to design an automatic dialect identification system in the three dialects of Ao.

3.1 Introduction

As Ao is an under-resourced language, there are no publicly accessible speech resources. As a result, an original speech database was collected and annotated particularly for this dissertation which serves as the basis of this research. The Changki Mongsen Chungli-Ao (CMC-Ao) speech database was collected from speakers of Mokokchung district, a district of Nagaland in the North-East of India. The Mokokchung district is the hometown of the Ao community. It comprises six major mountain ranges, namely, Asetkong, Jangpetkong, Japukong, Langpangkong, Ongpangkong, and Tsurangkong ranges.

For this research, the speech data were collected from speakers of the Chungli dialect as spoken in Mopungchuket village, Mongsen dialect as spoken in Khensa village, and Changki dialect as spoken in Changki village. Mopungchuket village is located in the Southern Asetkong range, Khensa village in the Southern Ongpangkong range, and Changki village in the Western Jangpetkong range. Figure 3-1 shows the three villages on the map of Nagaland, India.

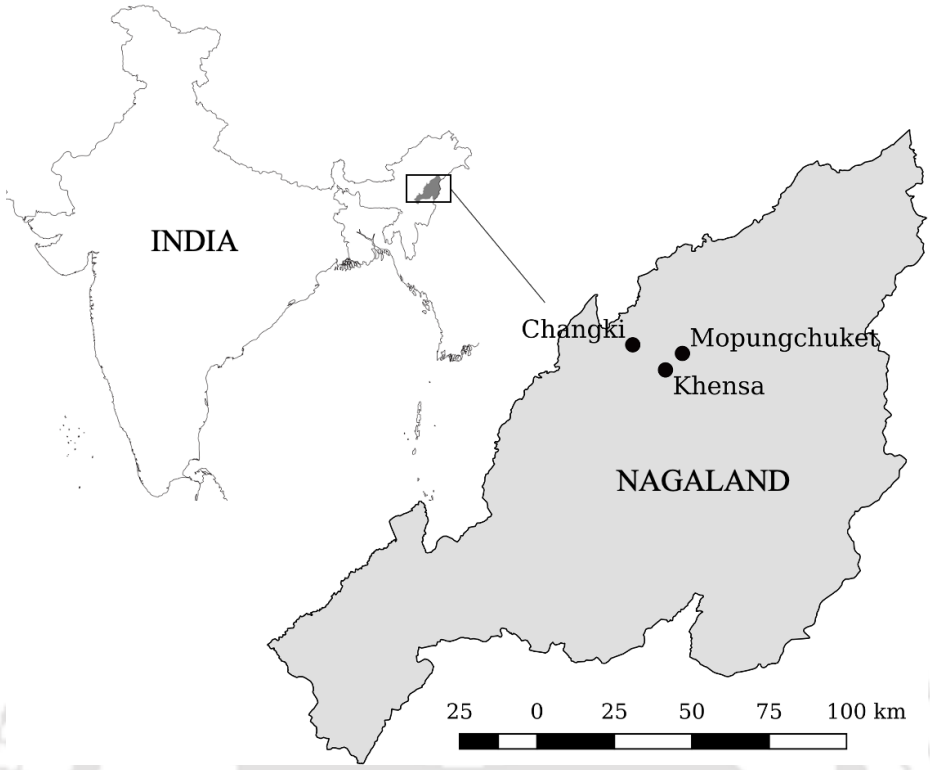


Figure 3-1: Data collection areas on a map of Nagaland.

3.2 Details of speech databases

This section describes the speech database that was explicitly collected for this dissertation. Three speech datasets were collected for each of the three dialects: disyllabic words (DiW), trisyllabic words (TriW), and passage level (PasL). The speech database materials are reported in Appendix A.

3.2.1 Disyllabic words (DiW)

The DiW corpus was recorded from 10 native speakers for each dialect in three contexts, namely, in a meaningful sentence, in isolation, and in a carrier phrase. The target words were six minimal sets (for each dialect) of disyllabic words contrasting in three lexical tones in Ao. Minimal sets are group of words that are same but differ in tones. One of the six sets, in Table 3.1, shows the tone assignment in the Ao disyllables in the three dialects. For the words in Table 3.1, the tone assignment in

the Changki dialect of Ao differs from that in the Mongsen and Chungli dialects for words with the same meaning. Table 3.1 shows that the Low tone of the Mongsen and Chungli dialects corresponds to the High tone of the Changki dialect, and the High tone of the Mongsen and Chungli dialects corresponds to the Low tone of the Changki dialect [2]. However, the tone change may not be consistent across the dialects. It is to be noted that the tones for the words 'rearing' and 'seeds', the HH tones terminate in glottal stops. In total, there were 20, 21, and 19 words across the six sets produced by the Changki, Mongsen, and Chungli speakers, respectively. Hence, in total, 600, 630, and 570 utterances were recorded from Changki, Mongsen, and Chungli dialects, respectively. The disyllabic utterances have an average duration of 0.45 sec. This dataset was collected specifically to investigate the acoustic features of the three tones in the three dialects. Hence, the words were not always the same across all dialects as the focus was to get sufficient examples of each tone. On the other hand, TriW corpus was constructed with words that are identical across the three dialects in terms of their segmental features, only differing in their tone assignment so that tone feature based DID can be attempted.

Table 3.1: Tone assignment in the word /metsü/ contrasting in tones across the three Ao dialects. The High, Mid, and Low tones of Ao are represented by H, M, and L.

	Changki	Mongsen	Chungli	Gloss
/metsü/	HH	MM	-	'deer'
	LH	HL	HL	'kick'
	-	-	HH	'rearing'
	LL	HH	HH	'salt'
	HL	MM	MM	'saliva'
	-	-	HH	'seeds'

An example of a Changki speaker producing the target word /metsü/ (salt) in three contexts is provided in 1-3 below.

(i) /ming ko metsü zükong/

'Put salt in the dish'

(ii) /metsü/

'salt'

(iii) /ni nü metsü te sano/

'I said salt'

3.2.2 Trisyllabic words (TriW)

The TriW corpus was recorded from 12 native speakers of each dialect, reading 40 trisyllabic words in three different contexts. These forty words are similar to a large extent across the three dialects in terms of their segmental features, however, they differ in tone assignment. These words were collected in order to see the dialect-specific information across the three dialects. As with DiW corpus, these 40 words were produced in three different contexts, namely, in a meaningful sentence, in isolation, and in a carrier phrase. This resulted in a total of 4320 utterances produced by speakers of all three dialects. The utterances have an average duration of 0.6 sec for the three dialects. Apart from that, in order to add session variability, the same speakers were recorded speaking the same utterances after two months of the first recording session. Hence, the total amount of data for TriW corpus was 8640 utterances across two sessions. For example, the three contexts of a Chungli speaker uttering the target word '/pilaba/' is presented as

(i) /Tsüngrem nungi pilaba ji nisung ka takum nung takoksa tulutiba ji lir/

'To part from God is the greatest loss in a man's life'

(ii) /pilaba/

'to part'

(iii) /ni nü pilaba ashi/

'I said to part'

3.2.3 Passage level (PasL)

The PasL corpus was recorded from 8 native speakers (4 males and 4 females) from each dialect. A brief narrative from the Bible, "The parable of the prodigal son" was read by each speaker. Among the three Ao dialects, Chungli is considered the standard variety of the language. In accordance with that, the literature is available only in the standard dialect as it is used in all formal gatherings and writings. Accordingly, the Bible passage was translated for the speakers of Changki and Mongsen dialects. The recording process was conducted for the speakers in four sessions for each dialect, reading the same passage to add session variability. Hence, the PasL speech database consisted of 96 passages with approximately 6 hours of recordings in total across the three dialects.

The CMC-Ao corpora can be summarised as follows:

- **DiW corpus**: disyllabic words, six minimal sets, 30 native speakers, 1800 utterances, read speech, three contexts
- **TriW corpus**: trisyllabic words, 40 words, 36 native speakers, 8640 utterances (across two sessions), read speech, three contexts
- **PasL corpus**: passage level, a Bible passage,"The parable of the prodigal son", 24 native speakers, 96 passages in total, read speech

3.2.4 Recording environment

Figure 3-2 depicts the general setup for recording the speech data, where the recordings took place in a real-world environment. Speech data was recorded with a TASCAM DR-100 MKII, a 2-channel portable digital recorder at a sampling rate of

44.1 kHz for all recordings. For high-quality recordings, the recorder was connected to a Shure SM10A head-mounted microphone.

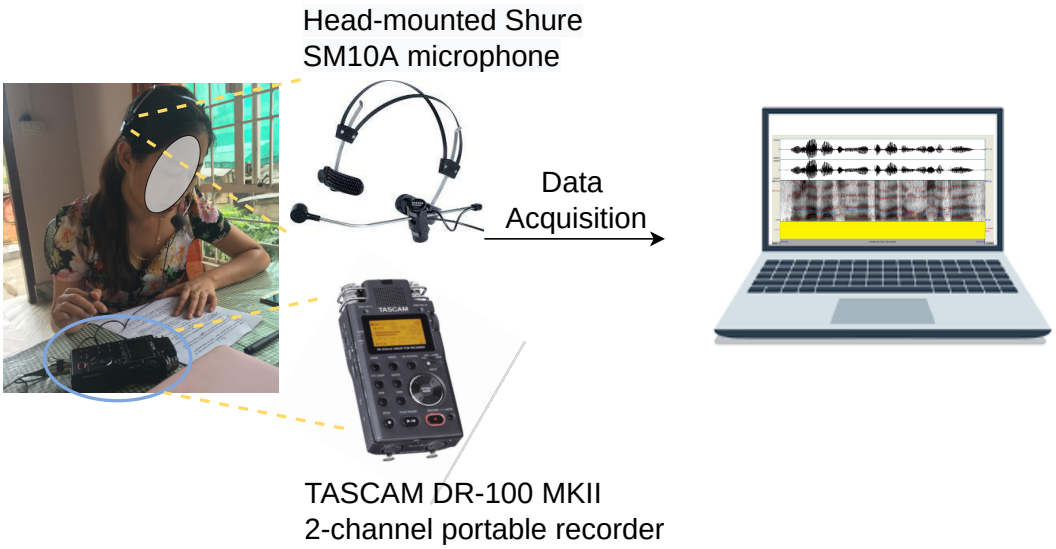


Figure 3-2: Demonstrating the overall setup for recording the Changki Mongsen Chungli-Ao (CMC-Ao) Corpora.

3.2.5 Annotation schemes

Once the recording was completed, speech data were transferred to a computer for annotation and archiving. The speech data (DiW and TriW corpora) were manually annotated using Praat 6.0.35 [68] for (a) target word boundary and (b) tone boundary, as shown in Figure 3-3. The annotation for tone categories was based on the tone specifications provided in the literature and from native speaker's judgement of the first author.

3.3 Speaker details

The number of speakers from the three dialects of Ao is shown in Table 3.2 based on the three types of corpora, namely, the DiW, TriW, and PasL corpora. There are

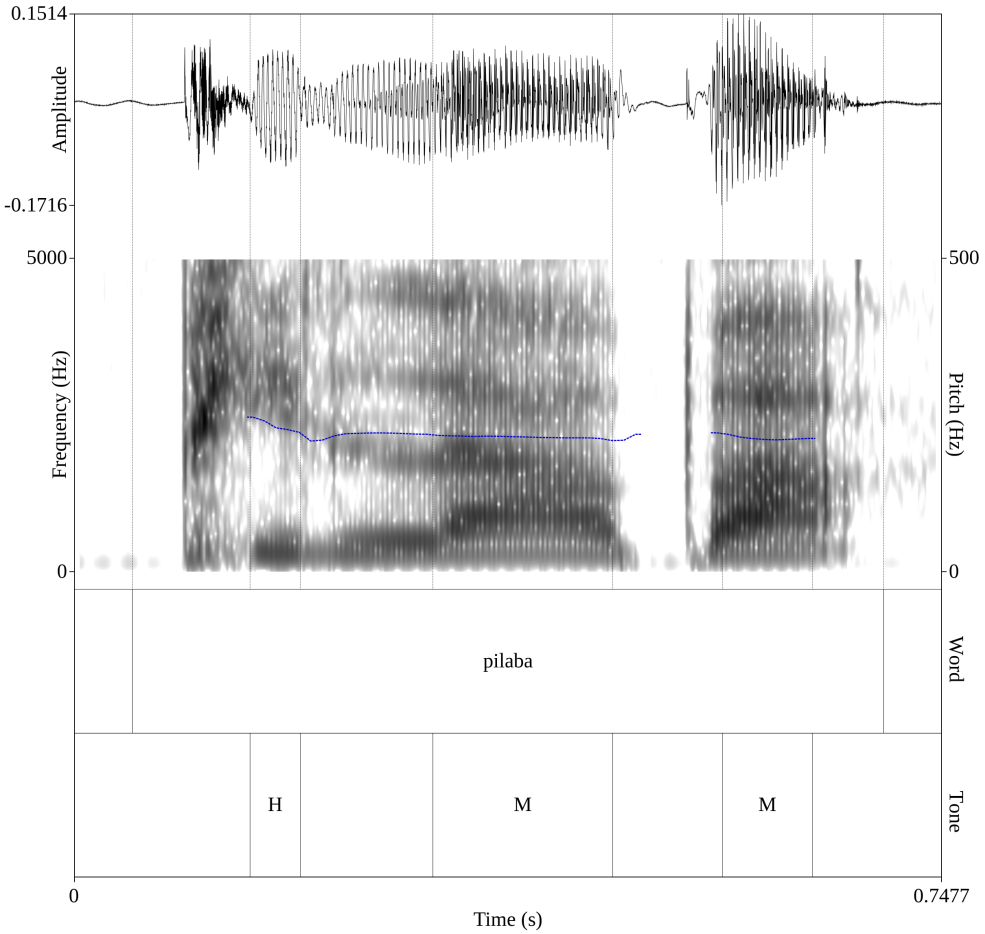


Figure 3-3: Example of a manual annotation.

4 Changki speakers (2 males and 2 females), 2 Mongsen speakers (2 females) and 8 Chungli speakers (4 males and 4 females) that participated in multiple corpus recordings in the database collection. All of the speakers recorded for this dissertation are either bilinguals, trilinguals, or multilingual between the ages of 24 and 60. Aside from their native Ao varieties, all of the speakers spoke English and Nagamese, a creolized variety of the Assamese language. The overall speaker details who participated in the field study for this dissertation are listed in Table B.1, Table B.2, and Table B.3 for DiW, TriW and PasL corpora, respectively of Appendix B.

Table 3.2: Dialect-wise distribution of participants for DiW, TriW, and PasL corpora based on gender.

	DiW corpus		TriW corpus		PaL corpus	
Dialect	Female	Male	Female	Male	Female	Male
Changki	8	2	6	6	4	4
Mongsen	5	5	6	6	4	4
Chungli	6	4	6	6	4	4
Total	30		36		24	

3.4 Summary

This chapter describes different categories of the Changki Mongsen Chungli-Ao (CMC-Ao) corpora. The Ao speech database was collected solely for the purpose of this dissertation as the resources were not available in the public domain. The CMC-Ao corpora consist of three different datasets, namely, disyllabic words (DiW), trisyllabic words (TriW), and passage level (PasL). The DiW corpus consists of 600, 630, and 570 utterances from Changki, Mongsen, and Chungli dialects, respectively. For the three dialects, the speech data was recorded from 30 native speakers. The disyllabic words have an average duration of 0.45 sec. In contrast, the TriW corpus contains a total of 8640 utterances for the three dialects collected from 36 native speakers. The trisyllabic utterances have an average duration of 0.6 sec. The PasL corpus, on the other hand, comprises 96 passages recorded from 24 native speakers. The total duration of the PasL corpus for the three dialects is approximately 6 hours. As a result, the DiW corpus was used in Ao for acoustic analysis. Meanwhile, the TriW and PasL corpora were used to develop an Ao DID system to identify the three dialects of Ao automatically.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 4

Tonal Feature Based Dialect Identification

Overview

This chapter comprises of an acoustic study conducted on the three tones in the three dialects of Ao using DiW corpus. It was found that the acoustic characteristics of the tones in the Changki dialect are markedly different from that of the Chungli and the Mongsen dialects. Additionally, the tonal assignments showed dialect-specific patterns. This work also demonstrated that Fundamental Frequency (F_0) features obtained using the Zero-Frequency Filtering (ZFF) approach outperform F_0 features derived by the Praat method based on autocorrelation in the Ao DID system. Hence, in the latter part of the chapter, F_0 features are used for automatic dialect identification in the Ao dialects using TriW corpus with the inclusion of Mel Frequency Cepstral Coefficients (MFCC) and Shifted Delta Cepstral coefficient (SDC) features using the Gaussian Mixture Model (GMM). It is confirmed that in both text-dependent and text-independent dialect identification, the F_0 features improve the accuracy of classification.

4.1 Introduction

From the literature on Ao tones described in section 1.2, we understand that the three tones in Ao are present in all three dialects. Secondly, the tonal specification of the same word may be different from dialect to dialect, as shown in Table 1.2. However, we also notice that there is no exhaustive acoustic study describing the characteristics of the three categories of tones across the Ao dialects. Despite of these descriptional gaps, text-independent automatic discrimination and identification of Ao dialects have been attempted in [69, 70], with the help of the descriptions provided on tone assignment in Ao dialects in *Coupe et al.* and *Temsunungsang et al.* [1, 2]. In *Tzudir et al.* [69], automatic discrimination of the two dialects of Ao, namely, Mongsen and Changki, was attempted using F_0 as the feature on eleven words for

each dialect. The results show that conventionally derived F_0 values are less robust than F_0 values determined by the Zero-Frequency Filtering (ZFF) approach, which yielded an accuracy of 69%. In *Tzudir et al.* [70], MFCC features were combined with F_0 features to perform DID in two dialects of Ao, viz., Mongsen, and Changki for forty words each, which gave an accuracy of 86.2%. In none of the two works, dialect-dependent F_0 differences in tone categories were investigated. Hence, this chapter has two parts; firstly, the acoustic characteristics of the three tones in the three Ao dialects are investigated to see the dialect-specific differences in them. Secondly, text-dependent and text-independent DID in three dialects of Ao is attempted based on tonal characteristics on the Ao dialects. In both parts of this work, we compare two different methods of F_0 tracking, namely, autocorrelation based Praat [68] F_0 tracking and ZFF [71]. In *Tzudir et al.* [69, 70], an Ao DID system was built only for two dialects with a limited dataset and only two features. This chapter is an extension of *Tzudir et al.* [69, 70], where identification of Ao dialects is attempted on three different dialects with the addition of SDC features.

The rest of the chapter is organized as follows: Section 4.2 gives a brief description of the speech corpus. Section 4.3 reports the results of the acoustic analysis of tones in the three dialects of Ao. Section 4.4 describes the DID experiments. Section 4.5 discusses the results, and finally, this chapter is summarized in section 4.6.

4.2 Speech corpus

Two speech datasets from the CMC-Ao database were used in this work for each of the three dialects, as described in chapter 3. The DiW corpus is used for acoustic studies on Ao tones as disyllabic minimal set predominates in the language. While the TriW corpus is utilized to attempt automatic dialect identification in the three Ao dialects.

4.3 Acoustic analysis of tones in Ao

In order to see the pitch contours of each tone in the dialects, F_0 values were extracted using two methods, viz., the inbuilt autocorrelation based pitch tracker in Praat [68] and ZFF [71]. The analysis of the F_0 contours of tones using the two methodologies are described in the following subsections.

4.3.1 F_0 extraction and analysis using Praat

The F_0 values using the autocorrelation method were automatically extracted with a script on Praat 6.0.35 at every 2% of the total duration of the tone [68]. The values were then exported to a spreadsheet for further analysis. However, the initial six and the final five data points were excluded from the analysis as Praat was unable to estimate the F_0 of several tokens at the tone boundaries. In order to normalize the F_0 values for speaker effects, z-score normalization was used [72]. To show the differences in the realization of tones in the three dialects, we consider the production of the disyllable /mətsə/ from DiW corpus. The production of /mətsə/ differs in terms of tones in the three Ao dialects as shown in Table 3.1 (see chapter 3). Figure 4-1 shows the normalized F_0 values averaged over the DiW database, plotted separately for each syllable using Praat and ZFF based pitch tracking. From Figure 4-1, we observe that the same word is associated with different tones in the three dialects. It is also observed that the Praat generated pitch contour shows less variance when compared to the ZFF computed pitch contour.

In this study, we compare the canonical pitch contours of the tones in all the three dialects of Ao. In order to derive the canonical pitch contours for the three tones in the three dialects of Ao, we consider the utterances with different tone specifications in DiW corpus. To describe the canonical pitch contours of the three tones, that is, High (H), Mid (M), and Low (L) in each dialect, the whole set of disyllabic target

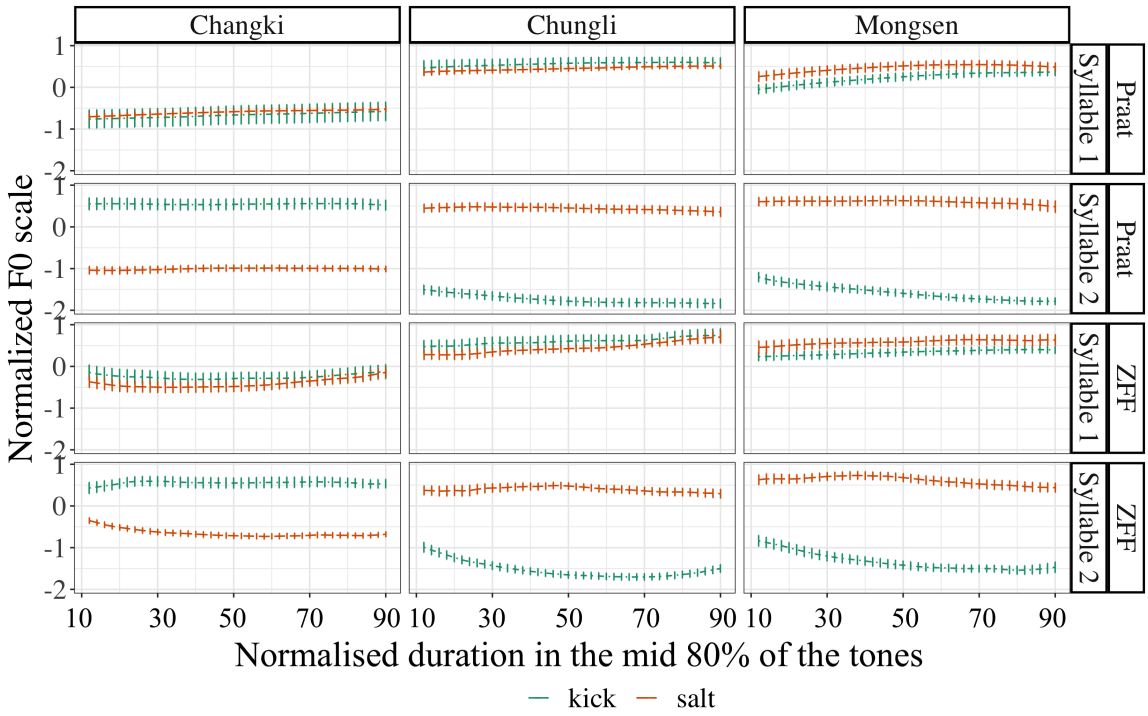


Figure 4-1: Normalized F_0 contours derived using Praat and ZFF for /mətsə/ in the three dialects of Ao.

words are considered from the three dialects. For the Chungli dialect, there are 570 utterances for six minimal sets (refer Appendix A) comprising of 1140 tokens of tones in total as the target words are disyllabic. Similarly, for the Mongsen dialect, there are 630 utterances for six minimal sets consisting of 1260 tokens of tones in total. Whereas, for the Changki dialect, it consists of 600 utterances for six minimal sets, with 1200 tokens of tones in total. In case of Praat derived F_0 values, due to short syllables and limitations of the pitch tracking algorithm, pitch was not detected in 368 tokens out of 3600 tokens. Hence, 368 tokens were not considered for the analysis. Figure 4-2 shows the z-score normalized F_0 contours with standard errors derived from Praat and ZFF for the tones in the three dialects. The plots show sufficient separation among the F_0 contours of the three categories of tones in the three dialects.

In order to see the dialect-wise differences in each category of tones, the three tones were plotted separately using Praat and ZFF F_0 values in Figure 4-3. The

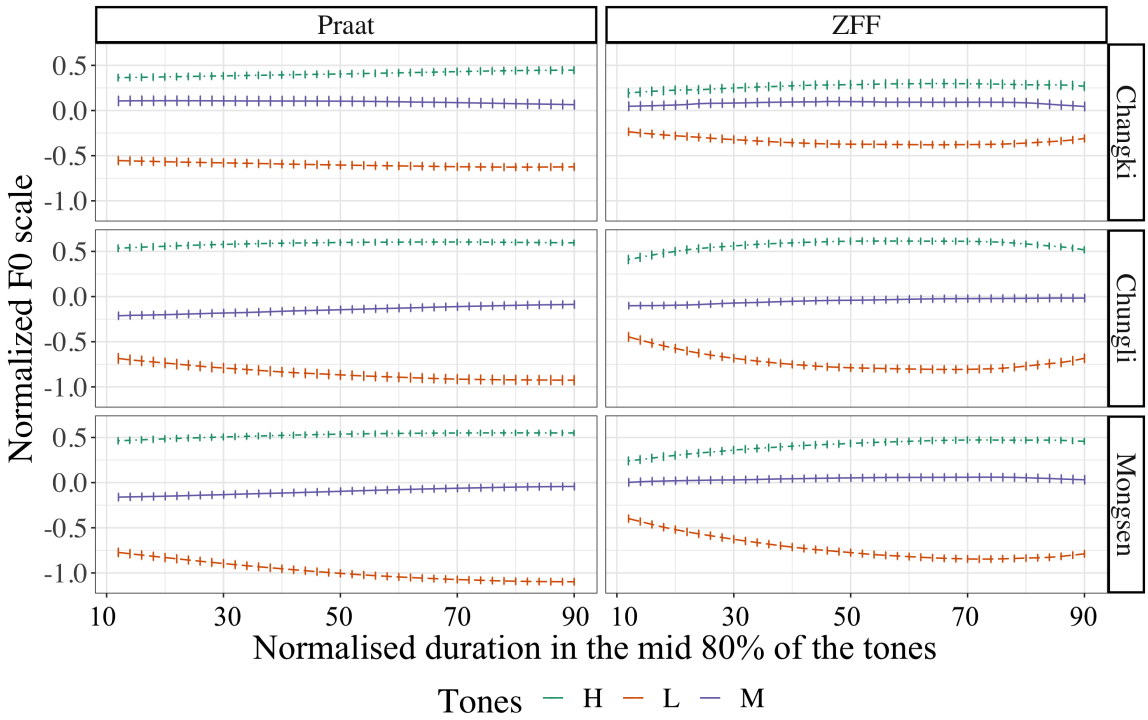


Figure 4-2: Canonical pitch contours of the three tones in Ao dialects derived using Praat and ZFF.

figure shows the relative differences among the same category of tones in the dialects of Ao. A visual examination of the tone contours in Figure 4-3 reveals that in terms of Praat derived F_0 contours, the high tone in the Changki dialect is lower than the Mongsen and Chungli dialects, and the low and the mid tones in Changki are higher than the other two dialects. While in ZFF derived F_0 contours, the low tone in the Changki dialect is higher than the Mongsen and Chungli dialects, and the high tone in Changki is lower than the other two dialects. A more detailed statistical analysis showing dialectal variation in the realization of tones in Ao dialects is presented in sub-section 4.3.3. We expect that these dialect-specific differences will also be able to aid automatic DID in the Ao dialects.

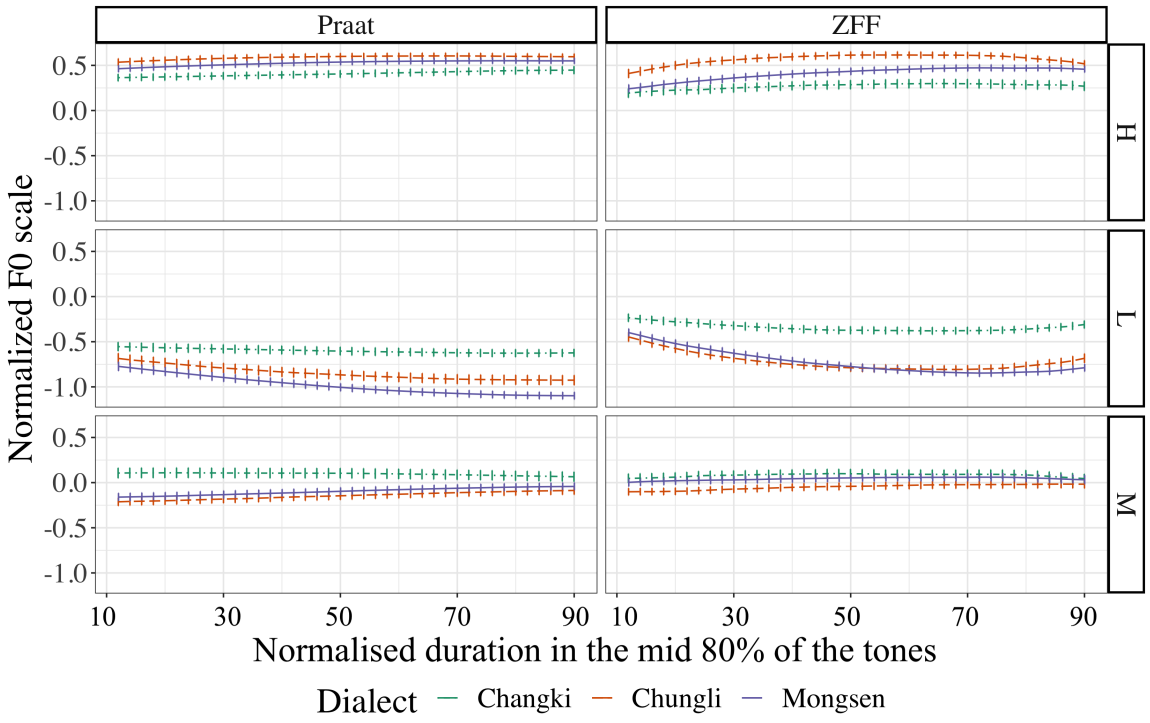


Figure 4-3: Normalized Praat and ZFF derived F_0 contours for the three tones in three Ao dialects.

4.3.2 F_0 extraction and analysis using Zero-Frequency Filtering (ZFF)

Murty et al. [73] proposed ZFF where the location of the epochs, i.e., the instants of significant excitation, gives an accurate estimation of the instantaneous Fundamental Frequency (F_0). The ZFF approach is computed to extract the epoch location from the speech signal where the Glottal Closure Instants (GCI) of the glottal cycles are assigned as epochs. In ZFF, the discontinuous impulse characteristics caused by the excitation source can be extracted by passing the speech signal through a cascade of two 0-Hz resonators in order to reduce the effects of high-frequency resonances. The steps involved in ZFF are as follows:

(i) To remove any time-varying low frequency in the signal, the difference of the

speech signal $s[n]$ is taken as

$$y[n] = s[n] - s[n-1] \tag{4.1}$$

(ii) The differenced speech signal $y[n]$ is then passed through a cascade of two ideal resonators at zero frequency which is given by

$$x[n] = -\sum_{k=1}^{4} a_k x[n-k] + y[n] \tag{4.2}$$

where $a_1 = 1$, $a_2 = 6$, $a_3 =$ -4 and $a_4 = 1$ [74].

(iii) Trend removal is then performed in $x[n]$ by subtracting the average value, which is computed over a window size, $(2N+1)$ at each samples resulting in the signal

$$\hat{x}[n] = x[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} x[n+m] \tag{4.3}$$

An auto-correlation function is used for computing the average pitch period of the windowed signal where the window size selection is dependent extensively on pitch period range. If the pitch period is in the range of one to two, window size selection is not critical. The resultant signal $\hat{x}[n]$ is sinusoidal, known as the ZFF signal, where the location of the GCIs are computed from positive zero crossings of the ZFF signal. The resultant instantaneous fundamental frequency (F_0) is computed by taking the inverse of the interval between two successive GCIs [71]. For the current study, the F_0 values were extracted from the Tone Bearing Units (TBU) using the ground truth and were linearly spaced to fifty values across the TBU. TBU are the vowel regions in a syllable associated with a tone. Z-score normalization was used to normalize the F_0 values [72]. In order to maintain uniformity between Praat and ZFF values, the middle forty data points were considered for analysis in ZFF. As discussed in sub-section 4.3.1, the ZFF derived pitch contour seems to capture more detailed

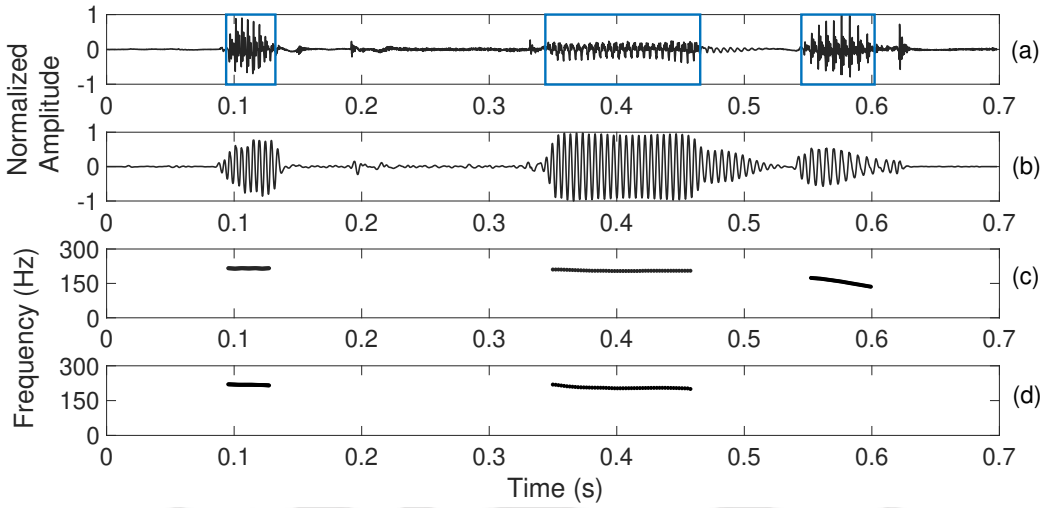


Figure 4-4: (a) Speech signal from a female speaker for the target word /baksaba/, (b) ZFF signal, (c) ZFF extracted pitch contour, (d) Praat extracted pitch contour.

pitch characteristics than the Praat derived pitch contours. However, ZFF could not estimate pitch in 22 tokens as they were extremely short and hence, excluded from the acoustic analysis.

From the discussions above, it is noticed that the Praat pitch tracker was unable to correctly estimate pitch at the tone boundaries of some of the tokens. For example, Figure 4-4 shows the plot for a female speaker. As seen in Figure 4-4 (d), Praat is unable to capture the pitch contour of the third syllable, while ZFF is as seen in Figure 4-4 (c). Hence, we decided to use pitch values extracted from ZFF in further analysis and classification of dialectal variation in Ao.

4.3.3 Statistical analysis

In order to confirm that the tone contours of the three tones are different and also to see if the tone contours differ across the three dialects, we built a Linear Mixed Effects (LME) model for average speaker normalized F_0 of the tone contour for each tone. The LME model is described by the distribution of two vector-valued random variables, namely, the response and the vector of random effects [75]. In the LME

Table 4.1: Results of the analysis of deviance test for the LME model with average F_0 as the dependent variable.

	Praat (N=3232)			ZFF (N=3578)		
Fixed Effects	df	χ^2	p-value	df	χ^2	p-value
Tone	2	1030.17	<0.001	2	711.86	<0.001
Dialect	2	8.13	<0.01	2	5.60	<0.05
Tone*Dialect	2	40.67	<0.001	2	68.10	<0.001

model, speaker normalized average F_0 was the dependent variable, and tone, dialect, and their interaction were the fixed effects. Speaker and tone location in the syllable (first or second syllable) were the random effects. The model was built using the *lme4* package on R [75, 76]. The model was further subjected to an analysis of deviance using Type II Wald χ^2 tests using the *car* package on R [77]. Pairwise comparisons between tones in the three dialects were conducted using the *emmeans* package on R. The results of the exploratory statistical tests are summarized in Table 4.1.

The results in Table 4.1 show that both tone category and dialect have significant effects on the average F_0 calculated using Praat and ZFF methods. The interaction of tone categories and dialect exhibited significance, confirming that realization of F_0 in the tone categories has dialect-specific differences.

Table 4.2: Results of pairwise comparisons of F_0 for each tone across the three dialects of Ao.

Method	Contrasts	High	Mid	Low
Praat (N=3232)	Changki-Chungli	n.s.	<0.01	n.s.
	Changki-Mongsen	n.s.	n.s.	<0.0001
	Chungli-Mongsen	n.s.	n.s.	n.s.
ZFF (N=3578)	Changki-Chungli	<0.001	n.s.	<0.0001
	Changki-Mongsen	n.s.	n.s.	<0.0001
	Chungli-Mongsen	n.s.	n.s.	n.s.

In order to investigate inter-dialect differences in average F_0 values, pairwise comparisons were conducted using the LME models, as shown in Table 4.2. The results from the pairwise comparisons show that in terms of average F_0 values in tone categories, the Chungli and the Mongsen dialects are quite similar to each other. However, the Changki dialect stands out as a distinct one from the other two dialects. Both in terms of ZFF and Praat derived F_0 values for the three tones, Chungli and Mongsen do not show any significant difference. However, ZFF and Praat derived F_0 contours show slightly different patterns in Changki-Chungli and Changki-Mongsen pairs. Figure 4-3 shows that in terms of Praat derived F_0 contours, the mid tone of Changki is the farthest from the mid tone of Chungli dialect. Hence, in Table 4.2 it is seen that the mid tones of Changki and Chungli are significantly different. Similarly, in terms of the low tone, the Mongsen and the Changki F_0 contours are the farthest apart, which is also supported by the pairwise comparisons in Table 4.2. While the overall pattern of the realization of F_0 contours for the tones in the two methods of F_0 extraction is similar, the magnitude of difference among the F_0 contours is different. For example, as seen in Figure 4-3, the difference between the low tone in the Changki dialect and the other two dialects increases when the F_0 contour is extracted with the ZFF method. Similarly, the difference between the high tone in Changki and Chungli increases in the ZFF derived F_0 contours. These differences are captured as significant in the pairwise comparisons shown in Table 4.2. While the dialectal differences are noticed in the realization of the tones, these differences are not consistent when different pitch extraction methods are used. Nevertheless, we expect that the dialect-specific differences in the realization of the tones may be captured in automatic DID systems.

4.4 Automatic DID in Ao

For the identification of dialects in Ao, three features are extracted, viz., ZFF [71], MFCC, and SDC. These features capture the micro-prosodic variations, spectral and temporal domain characteristics of the signal. As this is a preliminary work, Gaussian Mixture Model (GMM) is used for DID as it can learn the data distribution even for limited speech datasets [78]. Also, the likelihood function in GMM is based on a well-understood statistical model and is insensitive to the temporal aspects of the speech.

In this work, ZFF derived F_0 was one of the features used for dialect classification. The F_0 values are extracted from forty linearly spaced points for every TBU. For example, in trisyllabic words, F_0 is a 3-dimensional vector with 1-dimension representing one TBU. The average F_0 of a TBU, calculated from the forty linearly spaced points, is used as a feature in dialect classification. Apart from that, ΔF_0 and $\Delta\Delta F_0$ features are also used to capture the differences in tones in the three dialects based on their slopes. The Δ features help to capture the rate of change of the F_0 contour. ΔF_0 is calculated by taking the successive difference between F_0 values, and $\Delta\Delta F_0$ is calculated by taking the successive difference between ΔF_0. The averaged F_0 features along with the change in F_0 values were arranged as follows:

- F_0: 3-dim feature vector
- ΔF_0: 3-dim feature vector
- $\Delta\Delta F_0$: 3-dim feature vector
- F_0+ΔF_0: 6-dim feature vector
- F_0+$\Delta\Delta F_0$: 6-dim feature vector
- ΔF_0+$\Delta\Delta F_0$: 6-dim feature vector

- F_0+ΔF_0+$\Delta\Delta F_0$: 9-dim feature vector

While DiW corpus was used only for acoustic analysis of Ao tones and to see if there is any dialect-specific acoustic differences, TriW corpus is used for automatic DID in Ao.

4.4.1 Baseline methods

Mel Frequency Cepstral Coefficient (MFCC)

Features extracted through MFCC prove to be one of the most successful features that capture the vocal tract information of speech [25, 37, 79, 35]. MFCC is derived from the nonlinear cepstral representation of sound since the human auditory system's perception of frequency components is on a logarithmic scale. As a result, a nonlinear Mel filter has been designed to emphasize the lower frequency components over the higher ones. In this process, MFCCs are extracted from the speech signal with a frame size of 20 ms and a shift of 10 ms. The cepstral coefficient is calculated by computing discrete cosine transformation of log filter-bank energies, which produces 13-dimensional feature vectors after normalized frame energy. The first and second-order derivative of the 13 coefficient is calculated. Hence, 39-dimensional feature vectors are extracted for each frame. An energy-based approach is used to detect the beginning and end points of the utterances. Cepstral Mean and Variance Normalization (CMVN) is performed after extracting the MFCC features [80].

Shifted Delta Cepstral (SDC)

SDC coefficients are useful as a feature vector for LID as it captures the speech temporal dynamics for a wide range of speech frames [81]. The method followed for LID is also followed in DID [37]. The SDC coefficients are based on four parameters, namely, $N_c - d - P - k$ where

- N_c is the number of cepstral coefficients computed for each frame $(c_0, c_1, ..., c_{N_c-1})$
- d represents the spread over which deltas are computed
- P determines the time shifts between successive delta computations
- k defines the number of delta-cepstral blocks whose delta-cepstral coefficients are stacked to form the final feature vector

Accordingly, for each SDC feature vector, kN_c parameters are used, as compared with $2N_c$ for conventional cepstra and delta-cepstra feature vectors. The 7-dimensional MFCC static coefficients $\{c_0, c_1....., c_6\}$ are appended with 49-dimensional dynamic features produced by applying a $7-1-3-7$ scheme resulting to 56-dimensional SDC features.

4.4.2 Identification of Ao dialects using GMM in TriW corpus

GMM is implemented for Ao DID, where a group of D^T dialects, $D^T = \{1, 2, 3\}$ is represented by GMM's λ_{1^T}, λ_{2^T}, λ_{3^T}. The objective is to find the dialect model which has the maximum *aposteriori* probability for a given observation sequence. Given a sequence of training vectors, maximum likelihood model parameters are estimated using the iterative Expectation–Maximization (EM) algorithm [82]. The feature vectors of L are assumed independent, so the maximum log-likelihood of a model λ for a sequence of feature vectors, $X = \{x_1, x_2, x_3...x_L\}$ [78], as stated in equation (4.4).

$$\hat{D}^T = arg \max_{1 \leqslant k \leq 3} \sum_{t=1}^{L} log\, p(x_t|\lambda_k) \tag{4.4}$$

Apart from the primary DID experiment, we also explored the effectiveness of tonal features using ZFF, and Praat estimated F_0 values. To compare the tonal features estimated using ZFF and Praat, TriW corpus is used, comprising of recordings

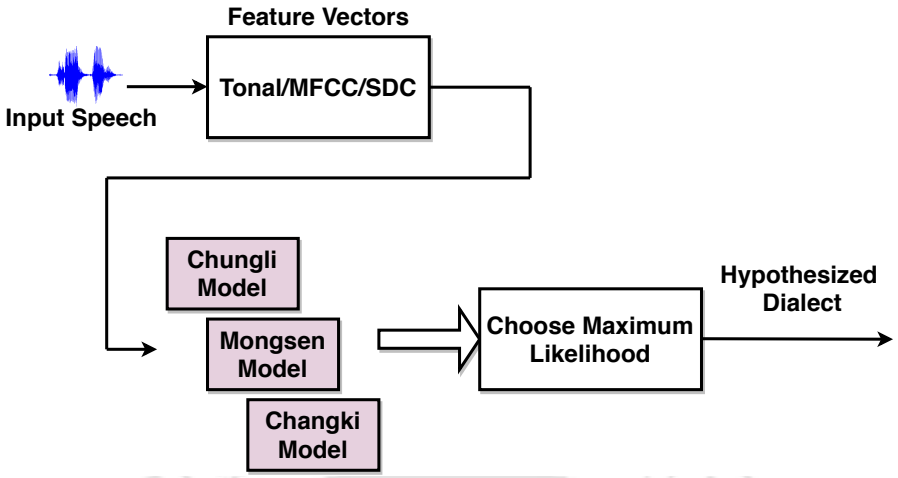


Figure 4-5: Block diagram of GMM based DID system.

from 12 native speakers of Ao. The training set consists of eight speakers (four male and four female) from their first session of recording, whereas the testing set consists of four speakers (two male and two female) from their second session of recording. As discussed in sub-section 3.2.2, the data consists of 1440 utterances for each dialect in the two sessions. In case of ZFF, the training and testing sets consist of 960 and 480 utterances, respectively. However, as Praat could not estimate F_0 values for 825 tokens, they were excluded from the experiment. In terms of Praat derived F_0 features, the training set consisted of 753, 715, and 884 utterances for Chungli, Mongsen, and Changki dialects, respectively. The testing set consisted of 359, 404, and 380 utterances for Chungli, Mongsen, and Changki dialects, respectively. Two separate models are built for each dialect using ZFF, and Praat estimated F_0 features, as shown in Figure 4-5.

For identifying the dialects in Ao, two scenarios are considered. In the first case (case-1), DID is text-dependent and in the second case (case-2) DID is text-independent. Initially, a DID system is developed using GMM based modeling approach. For both case-1 and case-2, TriW corpus is used, comprising of 12 native speakers for each dialect (six male and six female). During training, data from eight speakers (four male and four female) is used, while, four speakers' (two male and two

female) data is used for testing in both cases. In case-1, forty models, one for each trisyllabic word, are built for each dialect. The overall GMM based DID system is shown in Figure 4-5, where dialect models are replaced by forty word models. DID accuracy was obtained from each of the forty models and averaged. The training and testing data consisted of 24 and 12 utterances, respectively, for each model. In case-2, a model is built for each dialect, as shown in Figure 4-5. The training and testing data consisted of 960 and 480 utterances for each model. In order to account for session variability, recordings from the first session were used in training, and the second session recordings were used in testing for both case-1 and case-2.

Considering their importance in DID, as discussed in section 4.3, tonal features are used for identifying the Ao dialects, along with MFCC and SDC features. MFCC and SDC features are designed to capture the spectral changes among the three dialects, whereas, tonal features are expected to capture dialect-specific tone information. In GMM, for training, the number of mixture components is empirically chosen for the experiment. In case-1, 8 mixtures for the tonal features and 16 mixtures for MFCC and SDC features are considered. Whereas, 32 mixtures for tonal features and 512 mixtures for MFCC and SDC features are considered for case-2. During testing, a test dialect is represented by a sequence of feature vectors, and the log-likelihood produced by the model is calculated using equation (4.4). The dialect is determined using the model with the highest log-likelihood after analyzing the results from each model for tonal, MFCC, and SDC features. Further, for better dialect modeling, the scores from MFCC, SDC, and tonal features are combined. The combination score S^T_{comb} is obtained by equation (4.5).

$$S^T_{comb} = \alpha_T S^T_{mfcc} + \beta_T S^T_{sdc} + (1 - \alpha_T - \beta_T) S^T_{f_0} \tag{4.5}$$

In equation (4.5), S^T_{mfcc}, S^T_{sdc} and $S^T_{f_0}$ denotes the scores obtained from MFCC, SDC,

and tonal features, respectively. In equation (4.5), the α_T and β_T weight values range from 0 to 1 and 0 to $1 - \alpha_T$ at an interval of 0.05, respectively. For one iteration of α_T, β_T iterates at an interval of 0.05 from 0 to $1 - \alpha_T$. The dialect is determined based on the combination score with the highest likelihood.

4.5 Results using TriW corpus

In this section, we conduct 3 experiments to see how MFCC, SDC, and F_0 features contribute toward the classification of the three dialects of Ao. In experiment-1, we use average F_0, ΔF_0 and $\Delta\Delta F_0$ of tones to classify the dialects in order to see the effectiveness of tonal information in DID. In this experiment, TriW corpus is used to conduct text-independent DID. Moreover, this experiment is performed with F_0 features estimated from both ZFF and Praat methods. experiment-2 is conducted with TriW corpus, and in this experiment MFCC, SDC, and F_0 features are used for a text-dependent DID system. experiment-3 is also conducted with TriW corpus, and in this experiment MFCC, SDC, and F_0 features are used to build a text-independent DID system. In all the three experiments, classification was performed using GMM as detailed in sub-section 4.4.2. The results obtained from the three experiments are summarized below.

The results of experiment-1 as shown in Table 4.3, confirm that the Praat derived average F_0 feature yields the best dialect classification accuracy. However, in case of ZFF derived F_0 values, the combination of average F_0 and ΔF_0 provides the best classification result. In both Praat derived and ZFF derived pitch values, the same set of F_0 features, namely, F_0, F_0+ΔF_0, F_0+$\Delta\Delta F_0$ and F_0+ΔF_0+$\Delta\Delta F_0$ yield the top four accuracies in experiment-1. This confirms the effectiveness of F_0 features in DID in Ao.

Results of experiment-2 are summarized in Table 4.4 and presented in Figure

Table 4.3: Results of experiment-1. Dialect classification performance in % using ZFF and Praat derived F_0 values. The best performance is represented in bold.

	ZFF				Praat			
Features	Chungli	Mongsen	Changki	Avg	Chungli	Mongsen	Changki	Avg
F_0	44.2	40.6	45.2	43.3	47.4	46.8	54.6	**49.6**
ΔF_0	34.8	32.5	41.0	36.1	21.8	31.2	35.4	29.5
$\Delta\Delta F_0$	33.3	36.7	37.9	36.0	27.6	36.1	41.2	35.0
F_0+ΔF_0	43.8	38.5	45.6	42.6	42.9	41.3	50.4	44.9
F_0+$\Delta\Delta F_0$	44.8	42.5	48.1	**45.1**	43.2	37.1	51.0	43.8
ΔF_0+$\Delta\Delta F_0$	35.8	37.3	40.4	37.9	27.4	29.0	42.3	32.9
F_0+ΔF_0+$\Delta\Delta F_0$	38.3	36.3	49.2	41.3	34.2	34.7	49.9	39.6

Table 4.4: Results of experiment-2. Ao dialect classification performance for features with the highest accuracies, recall, precision, and F1-scores in %. The best performance is highlighted and represented in bold.

Features	Accuracy	Recall	Precision	F1-score
F_0	52.5	52.4	54.2	76.3
F_0+ΔF_0	54.2	53.4	57.0	77.1
F_0+$\Delta\Delta F_0$	51.7	51.7	54.3	75.8
MFCC (M)	73.1	72.7	75.5	86.5
SDC (S)	71.3	70.8	72.8	85.7
M+S	82.2	82.0	83.5	91.1
M+S+F_0	87.3	87.1	88.4	93.6
M+S+F_0+ΔF_0	**87.8**	**87.7**	**88.9**	**93.9**
M+S+F_0+ΔF_0+$\Delta\Delta F_0$	86.5	86.3	87.5	93.2

4-6. From Table 4.4, it is noticed that in terms of F_0 features, the best accuracy in discriminating the three dialects of Ao is achieved when F_0 and ΔF_0 features are combined. The average accuracy across the three dialects is 54.2%. When MFCC and SDC features are used in classification, the accuracy of Ao DID is 82.2%. However, when combined with a combination of F_0 features, namely, F_0+ΔF_0, the accuracy is the best, i.e., 87.8%. Hence, we confirm that addition of F_0 features improve dialect

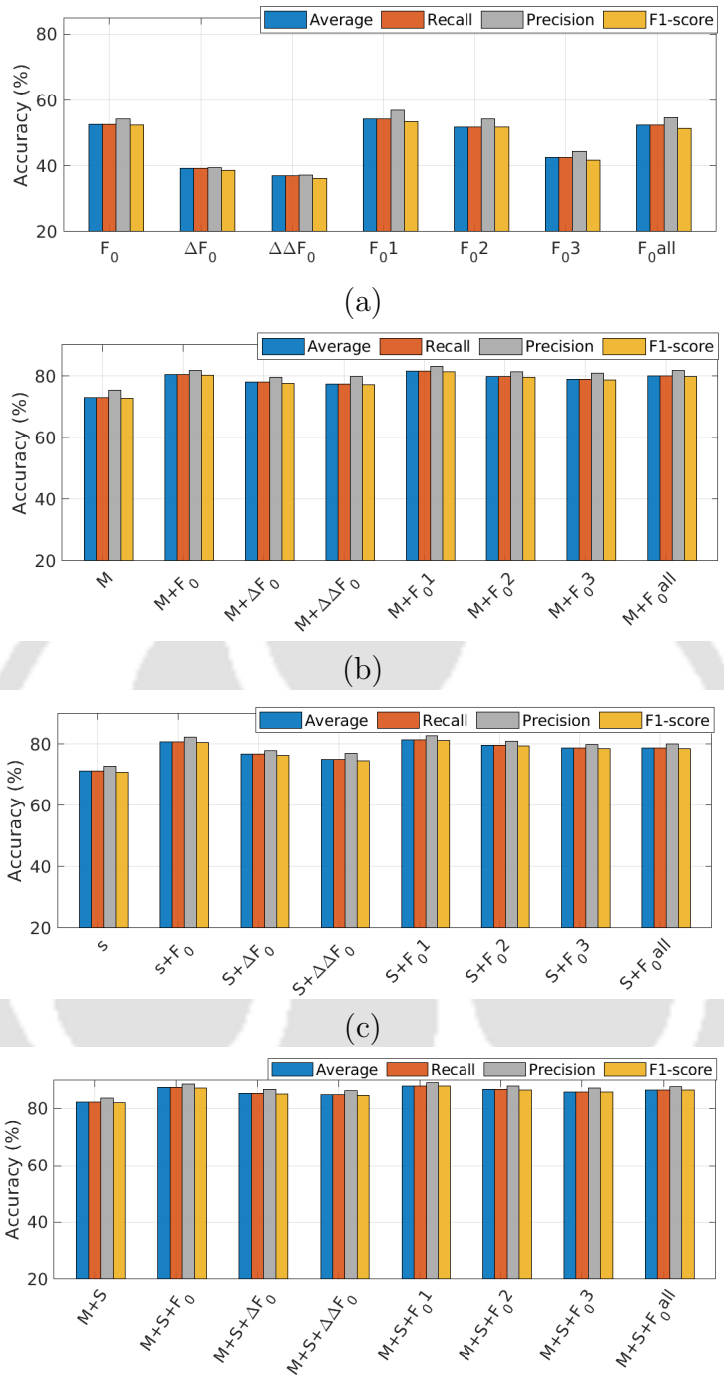


Figure 4-6: Results of experiment-2. Ao dialect classification performance with average accuracy, recall, precision, and F1 score. Abbreviations for the features and their combinations: F_0+ΔF_0 ($F_0$1), F_0+$\Delta\Delta F_0$ ($F_0$2), ΔF_0+$\Delta\Delta F_0$ ($F_0$3), F_0+ΔF_0+$\Delta\Delta F_0$ (F_0all), MFCC (M), SDC (S) etc.

Table 4.5: Results of experiment-3. Ao dialect classification with the features with the highest accuracies, recall, precision, and F1-scores in %. The best performance is highlighted and represented in bold.

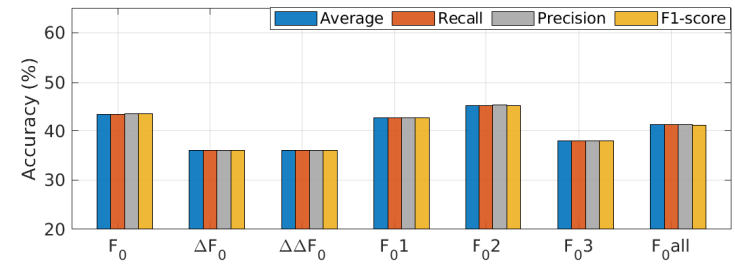
Features	Accuracy	Recall	Precision	F1-score
F_0	43.3	43.3	43.5	43.4
F_0+ΔF_0	42.6	42.6	42.7	42.6
F_0+$\Delta\Delta F_0$	45.1	45.1	45.2	45.1
MFCC (M)	62.8	62.8	64.2	62.9
SDC (S)	64.8	64.8	65.3	64.7
M+S	67.9	67.9	69.0	68.0
M+S+F_0	**70.4**	**70.4**	**71.2**	**70.5**
M+S+F_0+ΔF_0	68.1	68.1	68.9	68.1
M+S+F_0+ΔF_0+$\Delta\Delta F_0$	69.0	68.0	68.9	68.0

classification in Ao.

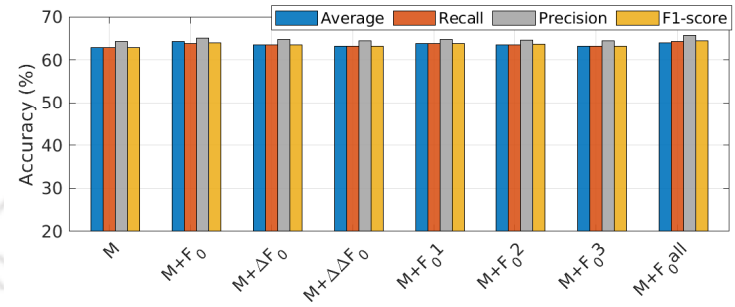
Results of experiment-3 are shown in Figure 4-7 and summarized in Table 4.5. The maximum average accuracy in terms of F_0 features is achieved when average F_0 is combined with $\Delta\Delta F_0$ as shown in Table 4.5. This combination yields an average accuracy of 45.1%. In Ao DID, when MFCC and SDC features are combined, the accuracy is 67.9%. The addition of the F_0 features improve the accuracy of DID as seen in the bottom three rows of Table 4.5. The best accuracy is achieved when MFCC, SDC, and average F_0 features are combined.

4.6 Discussion and summary

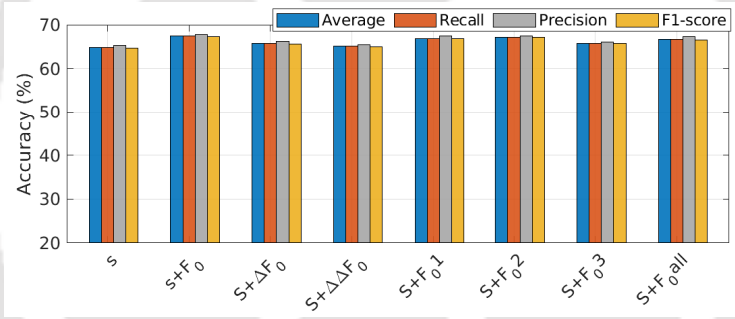
The acoustic analysis of the tones in the three dialects of Ao in this study showed that the acoustic feature of tones in the Changki dialect are distinct from the other two dialects, namely, Mongsen and Chungli. Apart from this, the assignment of tones also showed dialect-specific pattern. These dialect-specific features are captured by MFCC features and automatic DID is further enhanced by the inclusion of F_0 and



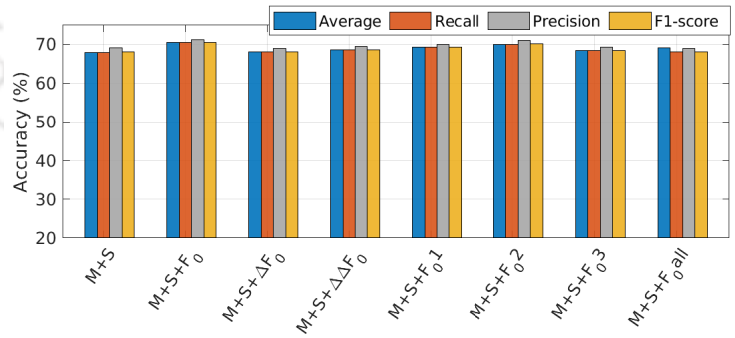
(a)



(b)



(c)



(d)

Figure 4-7: Results of experiment-3. Ao dialect classification performance with average accuracy, recall, precision, and F1 score. Abbreviations for the features and their combinations: F_0+ΔF_0 ($F_0$1), F_0+$\Delta\Delta F_0$ ($F_0$2), ΔF_0+$\Delta\Delta F_0$ ($F_0$3), F_0+ΔF_0+$\Delta\Delta F_0$ (F_0all), MFCC (M), SDC (S) etc.

SDC features. The F_0 feature is able to capture the dialect-specific tonal information, and the SDC feature captures the dialect-specific spectro-temporal information.

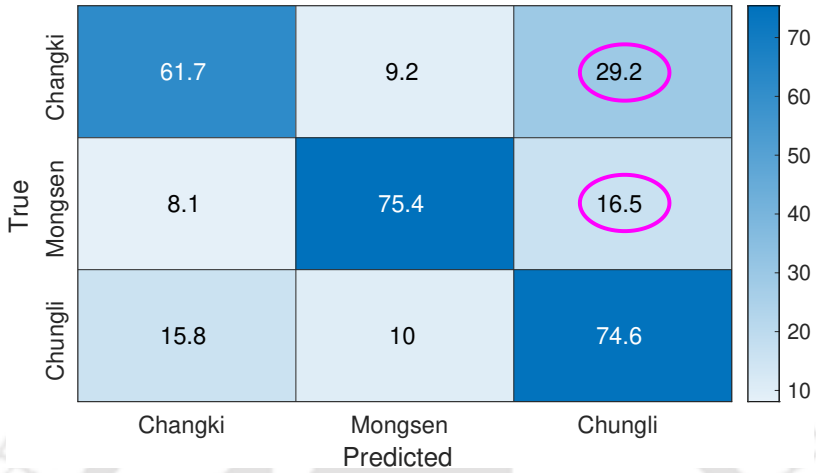


Figure 4-8: Confusion Matrix for the best feature combination (M+S+F_0) in % of experiment-3.

This work also showed that F_0 features derived by the ZFF method perform better in the Ao DID system when compared to a system where F_0 features are derived by autocorrelation based Praat method. As the ZFF method estimates F_0 from the GCI locations, consonant or phonation induced perturbations cannot affect F_0 estimation. Hence, the ZFF method is able to capture the F_0 contour reliably and with sufficient micro-prosodic information. These micro-prosodic information are dialect-specific which aid automatic DID. Apart from that, statistical analysis reported in Table 4.2 also substantiate that ZFF derived F_0 features are better predictors of Ao dialects.

In this study, a DID task performed on a text-independent speech corpus with F_0, MFCC and SDC features was able to yield an accuracy of 70.4% in classifying the three Ao dialects. When we look at the confusion matrix as in Figure 4-8, a systematic pattern emerged where both the Changki and the Mongsen dialects are misclassified as Chungli (circled in magenta). This is possibly due to the fact that in formal occasions the Changki and the Mongsen speakers tend to switch to the standard dialect, i.e., Chungli. Moreover, the written form of the language is based

on the Chungli dialect which might have biased the pronunciation of some speakers towards the Chungli dialect while reading the text in this study.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 5

Excitation Source Feature Based Dialect Identification

Overview

This chapter reports the findings of an automatic dialect identification task conducted on Ao speech data using excitation source features. In this direction, Residual Mel Frequency Cepstral Coefficient (RMFCC) feature is investigated to discriminate the three Ao dialects using Gaussian Mixture Model (GMM) in TriW corpus. Mel Frequency Cepstral Coefficients (MFCC) and Shifted Delta Cepstral (SDC) coefficients are used as the baseline features. The performance of the system yields a better dialect identification accuracy rate when all three features are combined. Further, Integrated Linear Prediction Residual Log Mel Spectrogram (ILPR-LMS) and gammatonegram of Linear Prediction (LP) residual, an approximate representation of source signal, are proposed for the dialect identification task. ILPR-LMS and LP-gammatonegram are used to exploit the time-frequency characteristics of the excitation source. This work proposes an attention-based CNN-BiGRU architecture for automatic DID tasks using PasL corpus. MFCC and LMS are used as the baseline features. As Ao is an under-resourced language, data augmentation was carried out to increase the size of the PasL corpus. The results showed that data augmentation improved dialect identification by 14%. *A perception test conducted on Ao speakers showed better dialect identification by the subjects when utterance duration was* 3 *sec. Accordingly, automatic dialect identification was conducted on utterances of various duration. A baseline dialect identification system with the* S_{lms} *feature attained an average* $F1$*-score of* 53.84% *in* 3 *sec long utterance. Inclusion of source features,* S_{ilpr} *and* S_{LP-gm}*, improved the* $F1$*-score to* 60.69%. *In a final system, with a combination of* S_{ilpr}*,* S_{LP-gm}*,* S_{lms} *and MFCC features, the* $F1$*-score increased to* 61.46%.

5.1 Introduction

This chapter reports the development of an automatic DID system in the Ao language, using different attributes of excitation source information. In the previous chapter, automatic DID in Ao incorporates the tonal information extracted from F_0 with the Vocal Tract (VT) representations [69, 70, 83]. The VT information is captured using the MFCC and SDC features. However, considering the role of F_0 in tone languages, we believe that excitation source information also plays a crucial role in distinguishing dialectal variations in such languages. On the contrary, to the best of our knowledge, excitation source information has not been studied in DID tasks. Therefore, this work is an attempt to explore the source features in distinguishing the Ao dialects.

In tonal languages, F_0 is an essential feature for DID systems [70]. Ao is a tone language where the assignment of tones is different across the dialects for the same word [2]. For example, in Table 5.1, the tone distribution is different across the three Ao dialects for the same word of the same meaning. However, it is reported that the tone alterations are not systematic across the three Ao dialects [83]. The differences in the assignment of tones across the three dialects motivated the present work to explore excitation characteristics for dialect identification in Ao.

Table 5.1: Tones in trisyllables across three Ao dialects.

Words	Changki	Mongsen	Chungli	Gloss
/pilaba/	HMM	MML	HML	'to part;to separate'
/temesen/	HMM	HHL	MHL	'liver'
/wamaba/	LHM	LHL	HHL	'to slice into pieces'

Excitation source features have been explored for LID tasks in clean and degraded environments [84, 28]. *Nandi et al.* [28] found that the characteristics of excitation and articulatory source are distinct for each sound unit. Thus, the characteristics of the same sound unit of a tonal language may differ due to the co-articulation effects

and the different assignment of lexical tones across the dialects. Hence, this work utilizes excitation source features such as the Residual Mel Frequency Cepstral Coefficient (RMFCC) and Integrated Linear Prediction Residual Log Mel Spectrogram (ILPR-LMS) [63, 64]. In addition, it is described in *Xu et al.* [65], recognition of tones is based on the production process and also on the human auditory perception of tones. Accordingly, this work is motivated to use the Gammatone filter as it is reported to effectively model human perceptual information [66]. In regard to the source information, LP residual is used as the input to generate gammatonegram from the response of the Gammatone filter, represented as LP-gammatonegram. Considering this, the gammatonegram of LP residual may carry dialect-specific information to distinguish the three Ao dialects, as it captures the human auditory response. Hence, it motivated us to build an automatic Ao DID system using the excitation source characteristics. The contributions of this chapter are listed below.

- RMFCC is investigated as a preliminary work using GMM to see the distribution of the three Ao dialects in DID task using TriW corpus.
- ILPR-LMS and LP-gammatonegram, a representative of the source signal, is explored next to learn the features automatically. The time-frequency representation is used to classify the Ao dialects.
- An attention-based Convolutional Neural Network-Bidirectional Gated Recurrent Unit (CNN-BiGRU) model is proposed. The proposed architecture utilizes the spatial (learned using CNN) and temporal context (learned using Bi-GRU) patterns of the spectrogram for the classification. Since Ao is a tonal language, the attention mechanism is used along the frequency direction. The attention mechanism provides higher weights to the frequency bins, which carry the discriminative patterns across dialects.
- The perceptual differences across dialects play a vital part in identifying dialects.

Hence, a perception test is performed with various segment duration.

The remainder of this chapter is organized as follows: Section 5.2 briefly describes the speech corpus recorded for this work. Section 5.3 describes the proposed work. The baseline methods are presented in section 5.4. Section 5.5 gives a description of the statistical analysis. The DID experimental setup and results are discussed in section 5.6 and section 5.7. Finally, section 5.8 summarizes the work.

5.2 Speech corpus

For each of the three dialects, two speech datasets from the CMC-Ao database were used in this study, as detailed in chapter 3. The TriW corpus is used to perform Ao DID task using GMM. While, attention-based CNN-BiGRU is used to automatically classify the three Ao dialects using the PasL corpus.

5.3 Proposed DID system for Ao language

This section describes the proposed work for the Ao DID system. Initially, the extraction process of RMFCC, ILPR-LMS, and LP-gammatonegram feature is described. Next, the proposed classifier architecture is discussed. In this work, the proposed features are extracted using a frame size of 25 ms and a shift of 10 ms with a 16 kHz sampling frequency.

5.3.1 Residual Mel Frequency Cepstral Coefficient (RMFCC)

Speech signal consists of excitation source information as it comprises the effect of vocal fold vibration, which modulates airflow through the glottis to produce sound. RMFCC is an excitation source feature that captures the spectral characteristics of the source signal. The RMFCC feature is computed from the Linear Prediction

(LP) residual spectrum in the cepstral domain. Non-uniform triangular Mel-filters are placed to process the logarithmic magnitude spectrum obtained from the LP residual. Discrete Cosine Transform (DCT) coefficients of the log magnitude spectrum are computed to get the RMFCC feature [85, 63]. Let $lpr[n]$ be the LP residual of the speech segment and $R[k]$ be the spectrum of $lpr[n]$. Let M_f be the Mel-filter-bank, which passes the log magnitude of $R[k]$ in order to convert it to the cepstral domain. The RMFCC feature, $RMFCC[k]$ is then evaluated as

$$RMFCC[k] = DCT[M_f(log|R[k]|)] \tag{5.1}$$

The 72-dimensional RMFCC feature vector for every speech signal utterances are considered, which includes the first 24-dimensions with their Δ and $\Delta\Delta$ derivatives. Figure 5-1 (g)-(i) illustrates the RMFCC features extracted for the vowel /a/ from the middle syllable in the target word /pilaba/ in the three dialects of Ao. The first 24 coefficients are plotted, excluding the zeroth coefficient. From Table 5.1, it is noticed that the tone for the middle syllable in the word /pilaba/ across the three dialects is Mid (M) tone. From Figure 5-1 (g)-(i), it is observed that for the same tone in the three Ao dialects, the trend is different for the RMFCC feature.

5.3.2 Integrated Linear Prediction Residual (ILPR)

The analysis and processing of speech signals are composed of separate representations for VT systems and excitation source. In Linear Prediction (LP) analysis, the LP coefficients (LPC) constitute the time-varying VT information, and the residual signal represents the excitation source characteristics [86]. Figure 5-2 shows the plot from female speakers of the three dialects. It represents a speech segment of 2 sec with lexical differences across the three dialects with the same meaning sentence. The plot represents the ILPR log spectrogram with distinct high variations of harmonics

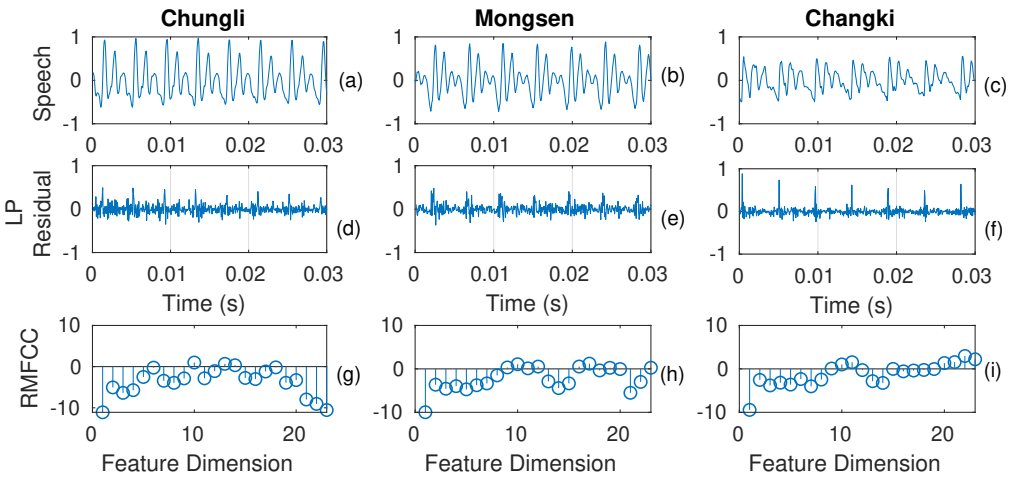


Figure 5-1: Illustration of excitation source characteristics across the three dialects in Ao for the vowel /a/. The speech signal is from a female speaker of each dialect from the middle syllable in the target word /pilaba/. Chungli: (a) Speech signal, (d) LP residual signal, and (g) RMFCC. Mongsen: (b) Speech signal, (e) LP residual signal, and (h) RMFCC. Changki: (c) Speech signal, (f) LP residual signal, and (i) RMFCC

in the Chungli dialect. As described in *Tzudir et al.* [83], the Chungli dialect has the highest H tone. Therefore, variations in tone and its harmonics are expected to be comparatively higher than the other two dialects. This might lead to more wavy patterns of tone and its harmonics, as shown in Figure 5-2. Likewise, the H tone of the Mongsen dialect lies in between the Chungli and Changki dialects [83]. Accordingly, the Mongsen dialect has lesser variations in its harmonics compared to the Chungli dialect. In contrast, the tonal space among the three tones in the Changki dialect is comparatively very less, as reported in *Tzudir et al.* [83]. Hence, the lower tonal variation in Changki leads to nearly straight harmonics compared to the other two dialects, as shown in Figure 5-2. Hence, ILPR signal is explored as an approximate representation of the voice source signal. The extraction of ILPR signal is briefly explained next.

First, the speech signal $s[n]$ is pre-emphasized in order to enhance the high-frequency components. This pre-emphasized speech $s_e[n]$ is then used to predict the LPC values using LP analysis. In order to obtain the ILPR signal, the original

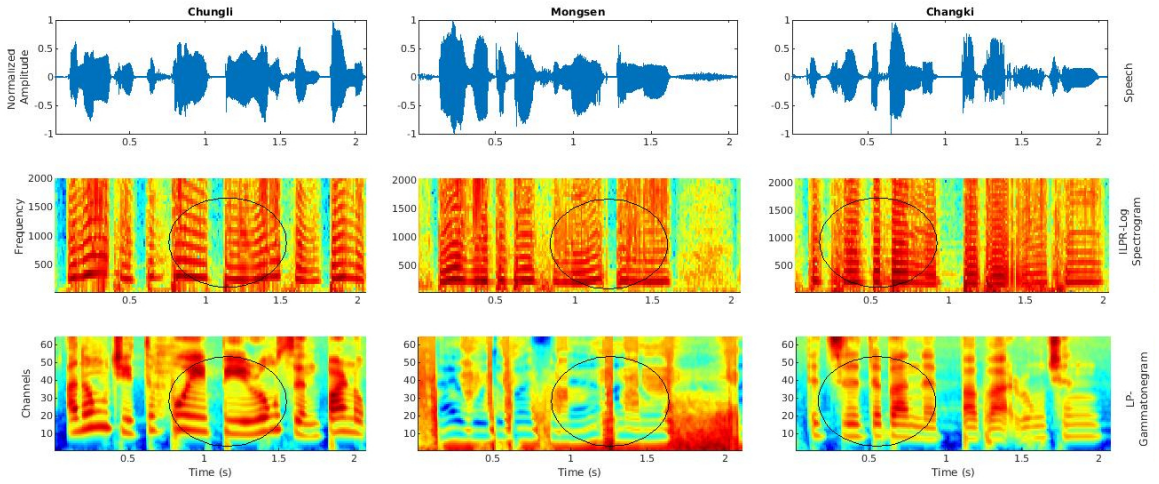


Figure 5-2: Different patterns present in the time-frequency representation of excitation source in the three dialects of Ao. First row: speech signal, Second row: ILPR-Log spectrogram, and Third row: LP-gammatonegram. The harmonic patterns present in the ILPR-Log spectrogram of Chungli resemble more wavy patterns with a higher dynamic range which reduces in the Mongsen dialect. The harmonic patterns tend to be straight in the case of the Changki dialect. The differences are circled. A similar pattern is observed in the LP-gammatonegram of the three dialects.

speech signal $s[n]$ is passed through the inverse filter obtained from LPCs [64]. The LP order p is set to $p = f_s + 4$, where f_s is the sampling frequency in kHz [64]. Further, the time-domain frame of the ILPR signal is transformed to frequency-domain by passing it through the Mel filter-bank. The Mel filter-bank consists of 40 overlapping triangular filters. Additionally, a natural logarithm is applied to the Mel spectrum. The Mel-spectra of consecutive frames are concatenated, and the resultant time-frequency representation, $S_{ilpr} \in \mathbb{R}^{40\times99}$, is obtained. Here, 40 represents feature dimension and 99 represents time-axis. Hence, the ILPR-LMS representation exploits the time-frequency characteristics of the excitation source signal for Ao DID task.

5.3.3 LP-gammatonegram

Ao being a tonal language, tone plays a vital role in classifying the dialects. As stated in the literature, both speech production and perception by humans have an impact on tone recognition [65]. The cochlea inside the human ear plays an important role in human perception. Based on the frequency of the incoming sound, the basilar membrane inside the cochlea vibrates and generates energy [87, 88]. Gammatone filter-bank mimics the response of the cochlea membrane and has an impulse response as the product of gamma function with sinusoidal tone centered at frequency f_c [87, 88]. The design of these filters is modeled to generate an auditory map like a spectrogram when there is vibration inside the cochlea, which is termed a gammatonegram. Motivated by the distinct patterns present in the ILPR log spectrogram across dialects in Figure 5-2, LP-gammatonegram is explored. The plot of LP-gammatonegram in Figure 5-2 also shows a similar contrast in the harmonics across the three dialects.

The primary difference of gammatonegram in terms of the traditionally used spectrogram is that the spectrogram processes the input signal via a bank of band-pass filters with the same bandwidth. Whereas, in the Gammatone filter-bank, the bandwidth of various band-pass filters is increased proportionally with respect to the central frequency [89]. A given frequency difference is perceived much stronger at low frequencies than at high frequencies. As tones are typically located at lower frequencies, it is an important property for tonal languages. Therefore, the essential lower frequencies are highlighted in the gammatonegram representation, giving more details than higher frequencies. Thus, the gammatonegram is more appropriate than the spectrograms.

In order to extract gammatonegram, initially, the speech signal $s[n]$ is converted to LP residual signal $lpr[n]$ in terms of source information. In LP analysis, the LPC constitutes the time-varying VT information, and the LP residual represents the excitation source [86]. The LP residual signal is initially determined by predicting

the VT information of the speech signal and then suppressing it using an inverse filter formulation [90]. The LP order (p) is set to $p = f_s + 4$, where f_s is the sampling frequency in kHz. Further, $lpr[n]$ is decomposed through a 64-channel Gammatone filter-bank, whose center frequencies (f_c^k) are equally spaced between 50 and 8000 Hz on the equivalent rectangular bandwidth (ERB) scale [91]. The bandwidth and impulse response of the filter is obtained by f_c^k and the decay factor by distributing in proportion to the ERB scale as

$$ERB_c^k = B_{min} + \frac{f_c^k}{Q_{ear}} \tag{5.2}$$

where, $Q_{ear} = 9.26449$ and $B_{min} = 24.7$ are known as Glasberg and Moore parameters, and f_c^k is the center frequency with channel $k = \{1, 2, 3 \ldots 64\}$ [92].

A Gammatone filter is defined by its time-domain impulse response, which is computed as

$$g[n, f_c^k] = \delta n^{\tau-1} e^{-2\pi b^k n} cos[2\pi f_c^k n + \theta] \tag{5.3}$$

where δ is the magnitude of response, τ is the filter order, and θ is the phase in radians. The constant b^k is the decay factor, which determines the duration of the impulse response and the bandwidth of the filter. For the fourth-order filter, b^k is computed as in equation 5.4 [87].

$$b^k = 1.09 \left(\frac{f_c^k}{Q_{ear}} + B_{min} \right) \tag{5.4}$$

Finally, the output gammatonegram at each channel is determined as the convolution of the LP residual of speech signal $lpr[n]$ with the impulse response of the Gammatone filter at that channel, as in equation 5.5.

$$y[n, f_c^k] = lpr_i[n] * g[n, f_c^k] \tag{5.5}$$

where $lpr_i[n]$ is the i^{th} frame of LP residual signal and $y[n, f_c^k]$ is in the output for every i^{th} frame and is further concatenated to form gammatonegram time-frequency representation. The natural logarithm is applied to the gammatonegram, and the resultant time-frequency representation is considered as the log gammatonegram of LP residual (S_{LP-gm}). Thus, S_{LP-gm} is proposed to capture the time-frequency characteristics of the excitation source signal for DID task in Ao.

5.3.4 Attention-based CNN-BiGRU classifier

The Ao dialect classification system is shown in Figure 5-3. The input speech signal is initially pre-processed by resampling to 16 kHz, implementing z-score normalization, and detecting the speech regions by removing the silence regions. Next, the pre-processed speech is used to extract the features. The extracted features are then fed to the attention-based CNN-BiGRU classifier for the three-class classification task.

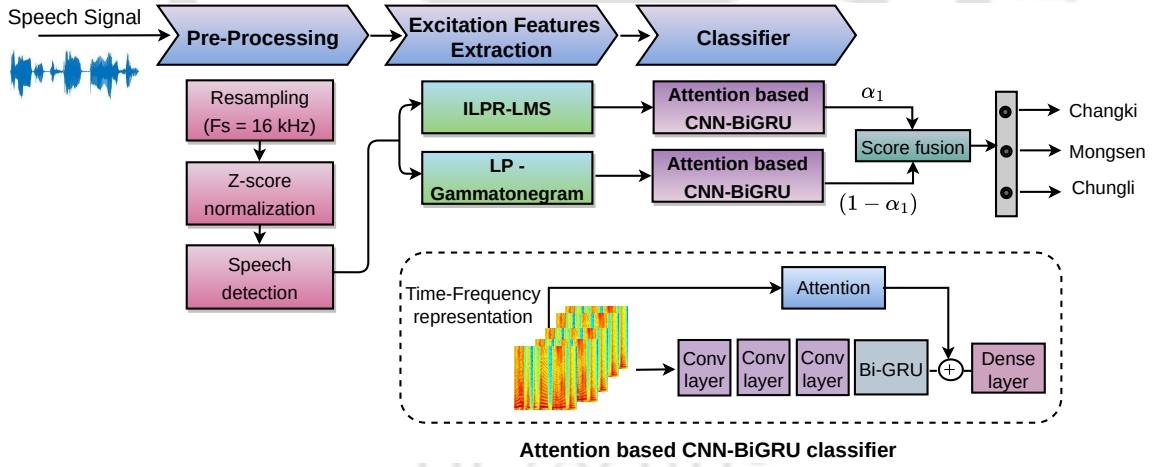


Figure 5-3: The proposed system for classifying three Ao dialects using an attention-based CNN-BiGRU classifier.

The architecture of the attention-based CNN-BiGRU Model is illustrated in Figure 5-4. attention-based CNN-BiGRU model is explored to learn the proposed features automatically. The proposed architecture learns spatial information of the spectrogram/gammatonegram using CNN architecture. The temporal context of different

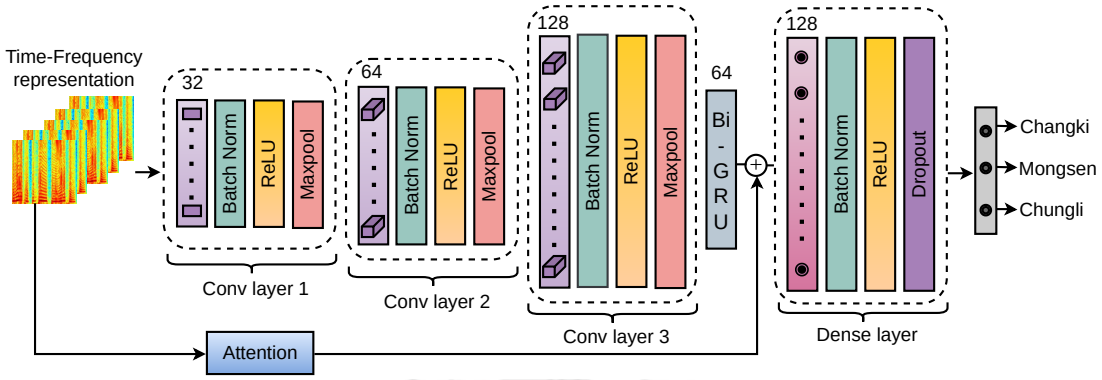


Figure 5-4: Attention-based CNN-BiGRU architecture.

dialects is captured by the Bi-GRU layer. Since Ao is a tonal language, the proposed model is motivated to use an attention mechanism [93] along frequency direction. This provides higher weights to those frequency bins, which have more discriminative information for classifying the dialects. Hence, the proposed architecture utilizes spatial, temporal, and frequency based attention for DID in Ao language. This work utilizes 1 sec segment duration ($\approx$ 99 frames) as the decision units for classification. For example, $S_{ilpr} \in \mathbb{R}^{40\times99}$ feature is passed as an input to the classifier. Here, 40 represents the feature dimension, and 99 denotes the time-axis.

The architecture consists of three 2D convolutional layers with 32, 64, and 128 number of kernels, respectively. A kernel size of $(3, 3)$ is used with a stride of $(2, 1)$. Batch normalization is performed to the output of each convolutional layer and is passed through a Maxpooling layer of pool size $(2, 1)$. The output of the last convolutional layer is fed to the Bi-GRU consisting of 64 units. Concurrently, the input gammatonegram is passed through the attention module, and its output is concatenated with the output of Bi-GRU. The concatenated output is passed through a 128 node fully-connected dense layer. Finally, the output of the dense layer is fed to the output layer (size=3) to predict the class label. The hidden layer uses $ReLU$ activation, while the output layer has $Softmax$ activation. The model is trained with a mini-batch size of 33 for 50 epochs. After the dense layer, a dropout of 0.4 is used.

The model is trained using categorical cross-entropy loss with the optimizer's initial learning rate set to 0.0001.

5.4 Baselines

This section describes the baseline features for this work. The baseline features are calculated using a window size of 25 ms and a shift of 10 ms with a 16 kHz sampling frequency.

5.4.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCC feature is commonly used in speech processing to represent the VT information [25, 37, 79, 35]. Accordingly, the pre-emphasized speech signal $s_e[n]$ is used to compute the MFCC features. For speech framing, a hamming window is used. This work utilizes 39-dimensional MFCC features (13+13Δ+13$\Delta\Delta$) for each frame [80]. CMVN is performed after extracting the MFCC features.

5.4.2 Shifted Delta Cepstral (SDC)

SDC coefficients are extensively used in the DID task as it represents the speech dynamics for a wider range of speech frames [37]. In order to obtain the SDC coefficients, it is based on four parameters, i.e., $N_c - d - P - k$ described in sub-section 4.4.1. The $7-1-3-7$ scheme is applied to the pre-emphasized speech signal $s_e[n]$ to obtain 49-dimensional features. This work uses 56-dimensional SDC features, which is derived by concatenating 7-dimensional MFCC static features with 49-dimensional dynamic features.

5.4.3 Log Mel Spectrogram (LMS)

The Mel spectrograms are obtained using a Mel filter-bank consisting of 40 overlapping triangular filters. The pre-emphasized speech signal $s_e[n]$ is used to extract the Mel-spectrograms. In addition, a natural logarithmic operation is performed on Mel-spectrograms, and the resultant time-frequency representation is considered as a log Mel spectrogram (S_{lms}). The $S_{lms} \in \mathbb{R}^{40\times 99}$ is also used as a baseline method in this work to provide complementary information with S_{ilpr} and S_{LP-gm}.

5.4.4 Gaussian Mixture Model (GMM)

GMM is used for Ao DID system, as it can learn the data distribution even for limited speech dataset [78]. Also, the likelihood function in GMM is computationally less expensive and is based on a well-defined statistical model. In addition, GMM is not sensitive to the temporal characteristics of speech. In this process, GMM model is created for each D^E dialects, $D^E = \{1, 2, 3\}$ and is depicted by GMM's λ_{1^E}, λ_{2^E}, λ_{3^E}. Considering a sequence of Z training vectors $X = \{x_1, x_2, x_3...x_Z\}$, the maximum log-likelihood $\hat{D}^E$ is computed as

$$\hat{D}^E = arg \max_{1\leqslant k\leq 3} \sum_{z=1}^{Z} log\, p(x_z|\lambda_k) \tag{5.6}$$

5.4.5 i-vector framework

The baseline method for Ao DID system in PasL corpus is developed using the i-vector framework [94]. Using a four-fold speaker-independent cross-validation training technique, the classification results are computed. Each iteration uses speech data from three folds for training and the remaining fold for testing with 1 male and 1 female speakers at every fold. Training is evaluated for every passage while testing is computed on segments of 1 sec duration. Using the train and test data, the 512

component gender independent Universal Background Model (UBM) is utilised to derive sufficient statistics. A total variability matrix (T-matrix) is estimated using the train data. Using a 150-dimensional T-matrix transformation, the i-vectors from the train and test data are extracted. Finally, cosine kernel scoring of the train and test i-vectors gives a score matrix for each of the three Ao dialects. The resultant dialect is determined based on the highest score.

5.5 Statistical analysis

Analysis of Variance (ANOVA) [95] is used extensively for the statistical test where it can assess one dependent variable at a time. However, to evaluate multiple dependent variables simultaneously, Multivariate Analysis of Variance (MANOVA) [96] is used. As the features are multi-dimensional, the MANOVA test is used to compute the statistical significance of the features. The three dialects are employed as the independent variables, and the features are considered as the dependent variables. The statistical significance across the dialects is examined using a statistical metric known as Wilk's lambda. It is a value that ranges from 0 to 1. A lower value of Wilk's lambda indicates that the feature is statistically significant.

5.5.1 Results on TriW corpus

Performance of statistical significance using MANOVA are presented in Table 5.2. Standalone features did not prove to be significant. However, when MFCC and SDC are fused with RMFCC, Wilk's lambda decreases. The lowest Wilk's lambda value is achieved for the combined features (MFCC+SDC+RMFCC). Hence, the combination of all three features is more significant for the current task.

Table 5.2: Performance of statistical significance analysis using MANOVA in trisyllabic words. Results are reported in terms of Wilk's lambda and p-value. The degree of freedom is represented by df. The best performance is highlighted and represented in bold.

Features	df	Wilk's lambda	p-value
RMFCC	2	0.94	<0.001
MFCC	2	0.89	<0.001
SDC	2	0.97	<0.001
MFCC+RMFCC	2	0.86	<0.001
SDC+RMFCC	2	0.89	<0.001
MFCC+SDC+RMFCC	2	**0.85**	**<0.001**

5.5.2 Results on PasL corpus

Statistical analysis is computed for the features obtained from the original speech data (without data augmentation). Table 5.3 reports the results of the MANOVA test across the individual and combined features. It is noticed that the p-value for individual and combined features is statistically significant. As seen in Table 5.3, the value of Wilk's lambda for the proposed feature S_{LP-gm} is lower than MFCC and S_{ilpr}. This shows that S_{LP-gm} contains class-specific information. This result encourages us to further explore them in combinations. As the features are combined, the value of Wilk's lambda decreases in comparison to the standalone feature. The lowest Wilk's lambda value is achieved when the source features are combined with the VT information ($S_{ilpr} + S_{LP-gm} + S_{lms}$). This signifies that there are dialect-specific differences in the features that can be used to model automatic DID systems.

5.6 Experimental setup

This section describes the various experimental setups used in this study. The section immediately following this describes the setup for Ao DID using GMM in TriW

Table 5.3: Performance of statistical significance analysis using MANOVA reported in terms of Wilk's lambda and p-value. The degree of freedom, ILPR-LMS, LP-gammatonegram, and LMS are represented by df, S_{ilpr}, S_{LP-gm}, and S_{lms}, respectively. The best performance is highlighted and represented in bold.

	Features	df	Wilk's lambda	p-value
Baselines	MFCC	2	0.87	<0.001
	S_{lms}	2	0.82	<0.001
	S_{lms}+MFCC	2	0.76	<0.001
Proposed	S_{ilpr}	2	0.89	<0.001
	S_{LP-gm}	2	0.83	<0.001
	S_{ilpr}+S_{LP-gm}	2	0.77	<0.001
	S_{ilpr}+S_{lms}	2	0.73	<0.001
	S_{LP-gm}+S_{lms}	2	0.69	<0.001
	S_{ilpr}+S_{LP-gm}+S_{lms}	2	0.64	<0.001
	S_{ilpr}+S_{LP-gm}+S_{lms}+MFCC	2	**0.59**	**<0.001**

corpus. The following sub-sections describe the classification using attention-based CNN-BiGRU in PasL corpus, data augmentation schemes, hyper-parameter tuning, and the use of various segment duration in the subsequent experiments.

5.6.1 Identification of Ao dialects using GMM in TriW corpus

GMM is used in Ao DID task to see the distribution of the three dialects. MFCC and SDC features are used as the baseline features. A model is trained for each dialect, considering the session and speaker variability. The trisyllabic data is divided into three-folds speaker-independent framework. There are twelve speakers in total for each dialect in the TriW corpus. The train set consists of four females and four males, comprising eight speakers in total from the first session recording for each dialect for each fold. The remaining four speakers with two females and two males are considered for the test set across the three dialects from the second session recording. Figure 5-5 illustrates the overall GMM based DID system where a model is generated for each

dialect. For every fold, the train set comprises 960 utterances, whereas the test set consists of 480 utterances for each model.

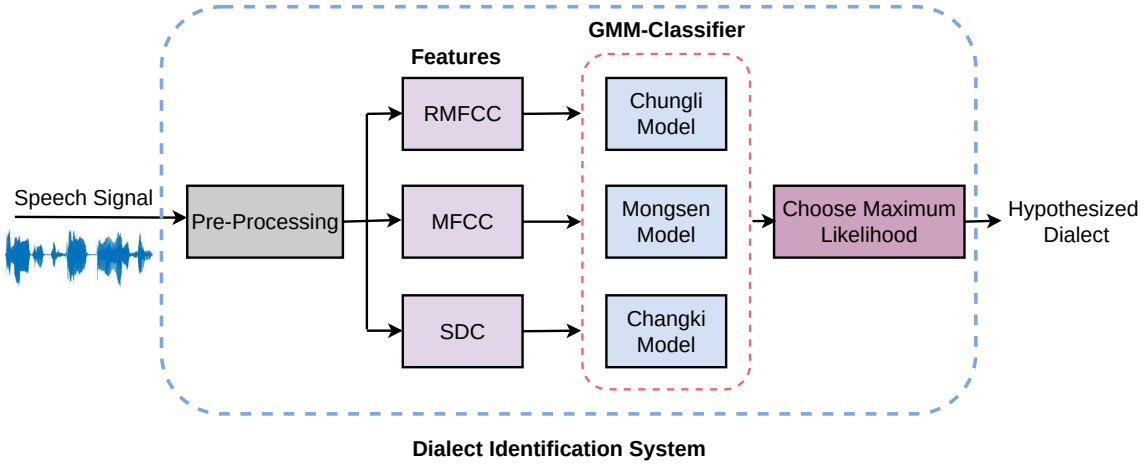


Figure 5-5: Block diagram of GMM based DID system.

For the experimental process, a 512 Gaussian mixture was chosen empirically for all three features. For testing, the sequence of feature vectors of the test dialect is fed to the models of each dialect to compute the log-likelihood. After evaluating the log-likelihood from each model, the dialect is identified based on the highest likelihood.

5.6.2 Classification using attention-based CNN-BiGRU in PasL corpus

MFCC and S_{lms} are considered as the baseline features. PasL corpus is used in this work where the speech data is divided into four non-overlapping folds in such a way that every fold contains the speech data of 1 male and 1 female speakers. At each iteration, the speech data from three folds are used for training and the remaining fold is used for testing. The training set is further divided into a 70 : 30 ratio to get the training and validation set, respectively. Therefore, four different sets of train, validation and test data are obtained from four-folds. All the folds contain different sets of speakers. Thus, the speakers of the test set are different from the train set speakers. Hence, the proposed approach is evaluated in a speaker-

independent framework. In order to utilize the temporal context of the speech data, the model is trained for 1 sec segment duration.

5.6.3 Training schemes

Data augmentation

In order to increase the dataset (PasL corpus) and avoid overfitting for the classification process, data augmentation is carried out in the following ways:

- Telephonic Speech:

  Speech data over telephone network can be used to augment the currently available dataset to lead to a better classification in training. Hence, the original PasL speech data was converted to telephonic speech using G.191, a software tool for speech and audio coding standardization [97]. The G.191 is a free software available on the International Telecommunication Union (ITU) website. In this work, the telephone quality speech signal is simulated using a pipeline process as suggested in [98]. Initially, the G.711 encoder is used to encode the G.712 filtered speech signal. Next, the G.711 encoded signal is encoded again and decoded with G.726 at 16 kbps. Again, the decoded signal is decoded again using the G.711 decoder. Finally, the decoded signal is filtered in the receiver direction as defined in ITU [99]. The speech data is downsampled to 8 kHz and simulated for telephonic speech.

- Reverberated Speech:

  The original PasL speech was augmented to two types of reverberated speech using the publicly available Roomsim toolbox [100]. The two types of reverberated speech differ in room sensor configurations and source configurations such as room size, sensor position, sensor direction, source orientation, source direction, etc. For instance, the first type of reverberated speech was configured

to a room size of 4.45 (in meters) with an omnidirectional sensor direction. In contrast, the second type of reverberated speech had a room size of 2.50 (in meters) with cardioid sensor direction.

After data augmentation, the PasL speech data resulted in $\approx$ 24 hours consisting of 384 passages across the three Ao dialects.

Hyper-parameter tuning

The parameters are tuned for the architecture described in sub-section 5.3.4. The 2D convolutional layer (C_1) of first Conv layer is tuned for 16, 32 and 64 number of kernels. The convolutional layer (C_2) of second Conv layer is tuned twice the number of kernels of C_1 and the convolutional layer (C_3) of third Conv layer is tuned four times the number of kernels of C_1. The kernel size and stride are kept fixed as $(3, 3)$ and $(2, 1)$. The number of Bi-GRU units is tuned for 64 and 128 units. Furthermore, the fully-connected layer is tuned for 32, 64 and 128 nodes. Figure 5-6 shows the performance of the average F1-score for all combinations of parameters where the experiment is conducted for one fold. It can be observed from the plot that the highest performance is achieved with C_1 = 32, Bi-GRU = 128, and DNN = 32. Hence, these parameters are utilized for the architecture for the subsequent experiments. The tuned architecture for Ao DID system is illustrated in Figure 5-7. The optimized architecture contains three convolutional layers consisting of Conv layers C_1, C_2 and C_3 with 32, 64 and 128 number of kernels, respectively.

5.6.4 Effect of segment duration

Dialect perception by Ao speakers

The perceptual difference across dialects is expected to be difficult even for humans to identify for a short segment of the speech signal. However, humans can easily

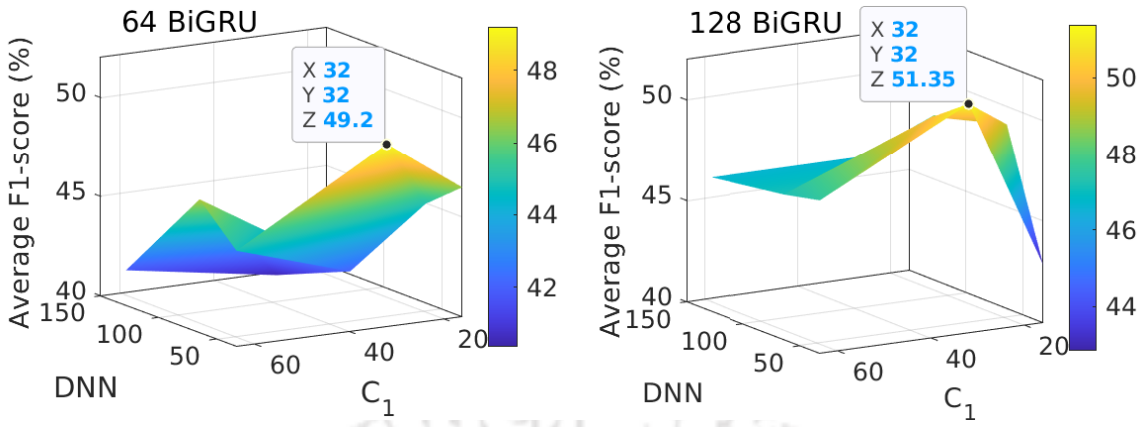


Figure 5-6: Illustrating performance for hyper-parameter tuning of the attention-based CNN-BiGRU classifier. The performance for 64 and 128 BiGRU units are demonstrated in the first and second sub-figures, respectively. X-axis represents the first Conv layer (C_1) tuned for 16, 32 and 64 number of kernels with C_2=2C_1 and C_3=4C_1. Y-axis represents the dense layers tuned for 32, 64, and 128 nodes. Z-axis represents the average F1-score in %.

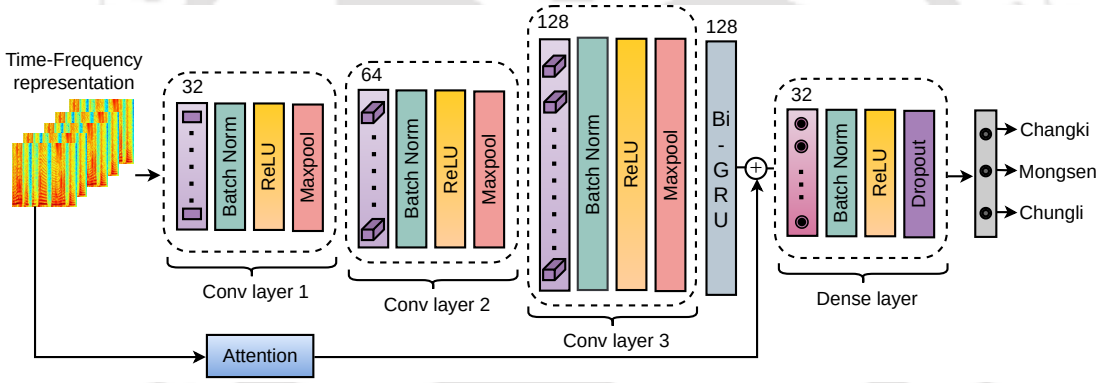


Figure 5-7: Optimized architecture after hyper-parameter tuning for Ao DID system.

identify the dialects if a sufficient duration of the speech signal is available. Therefore, a subjective evaluation of the speech data was performed to see the effect of segment duration across the three Ao dialects. The experiment was conducted on 15 subjects (native speakers) in total, comprising 5 subjects from each dialect. The list of subjects who participated in the perception test are presented in Table B.4 of Appendix A. The stimuli consisted of 60 speech chunks, 10 files from each dialect of 1 sec, and 3 sec segment duration. The speech data (PasL corpus) were extracted from the recorded passages described in section 5.2. The subjects from each dialect were given four

choices to choose from, namely, 'Changki,' 'Mongsen,' 'Chungli,' and 'Can't Decide.' The stimuli were presented randomly to the subjects, and they were asked to listen and identify the dialect they heard. Each subject took about 15 minutes to complete the task.

Automatic dialect classification

In the automatic method, the models are trained and tested on 1, 2, 3, 4, 5, and 6 second durations. The method to evaluate the effect of segment duration is the same as in sub-section 5.6.2. A four-fold speaker-independent cross-validation training technique is used to obtain the classification results.

5.7 Results

This section reports the results of Ao DID using GMM setup followed by classification results using attention-based CNN-BiGRU. This section also discusses the classification results obtained after data augmentation, classification performance with the optimized architecture, and finally with the use of speech data segments of various duration.

5.7.1 Identification results of Ao dialects using GMM

The classification performance of the Ao DID system using the GMM classifier is shown in Table 5.4 evaluated in three-fold speaker-independent training scheme. The excitation source feature, i.e., RMFCC, gives a decent performance of 47.87% in classifying the three Ao dialects. This result motivated us to fuse the source feature (RMFCC) with the VT features. As shown in Table 5.4, two feature combination (MFCC+RMFCC and SDC+MFCC) shows improvement in comparison with the standalone VT features.

Table 5.4: Dialect identification accuracies in % with 2 males and 2 females as the test set. The results are reported in terms of mean (μ) and standard deviation (σ) of three-fold cross-validation. The best performance is highlighted and represented in bold.

Features	Accuracy			
	Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
RMFCC	55.83±9.20	36.67±8.28	51.11±9.90	47.87±9.13
MFCC	31.18±8.39	66.11±0.99	68.54±3.92	55.28±4.44
SDC	64.24±6.44	55.28±4.31	75.63±2.74	65.05±4.50
MFCC+RMFCC	55.83±3.58	67.36±0.79	72.08±3.75	65.09±2.71
SDC+RMFCC	69.18±7.58	57.92±4.44	76.87±2.80	67.99±4.94
MFCC+SDC+RMFCC	**69.51±12.47**	**72.08±5.66**	**81.32±5.27**	**74.31±7.80**

Additionally, score fusion is conducted across the three features to improve the classification performance. The combination score S^E_{comb} is obtained by equation 5.7.

$$S^E_{comb} = \alpha_E S^E_{mfcc} + \beta_E S^E_{sdc} + (1 - \alpha_E - \beta_E) S^E_{rmfcc} \tag{5.7}$$

where, S^E_{mfcc}, S^E_{sdc} and S^E_{rmfcc} are the scores obtained from MFCC, SDC, and RMFCC features, respectively. The value of α_E varies from 0 to 1 with an interval of 0.05. The value of β_E varies from 0 to $1 - \alpha_E$ for each α_E iteration with an interval of 0.05. Hence, the dialect is identified depending on the highest likelihood combination score. For instance, Figure 5-8 shows the accuracy in percentage of one-fold for all iterations of α_E and β_E in the Chungli dialect. The highest accuracy is achieved for $\alpha_E = 0.3$ and $\beta_E = 0.5$.

As seen in Table 5.4, when the source feature is fused with MFCC and SDC combination (MFCC+SDC), notable improvement in the performance is seen. This shows the significance of the excitation source feature in discriminating the three dialects of Ao. Additionally, average F1-score, precision and recall are reported in

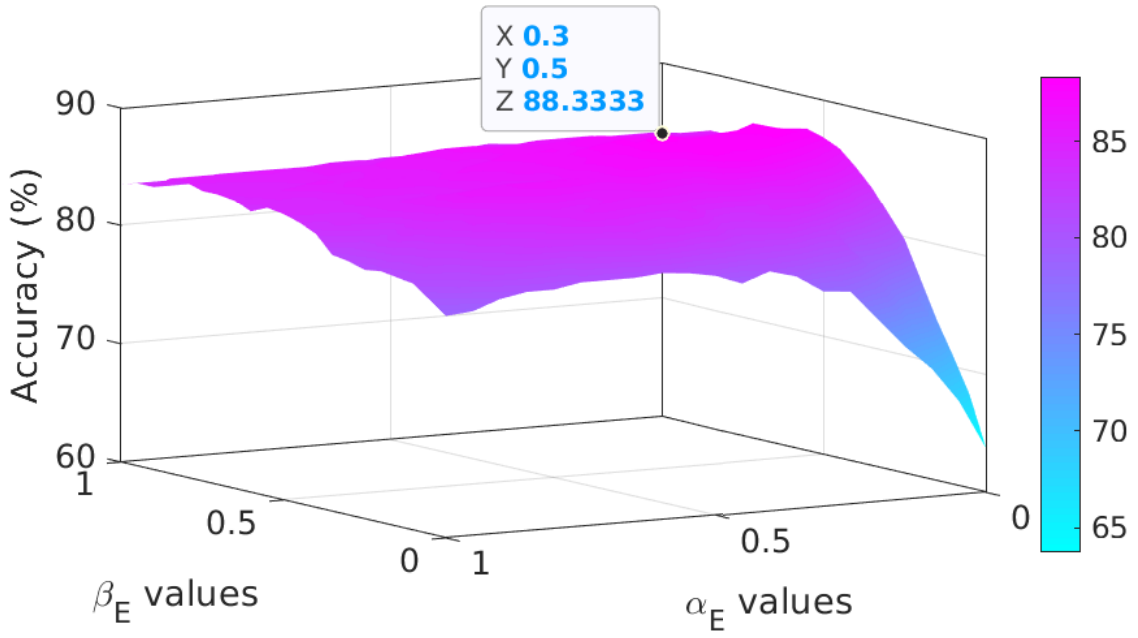


Figure 5-8: Accuracy plot in percentage for one-fold (out of three-fold) with respect to α_E and β_E in the Chungli dialect.

Table 5.5: Dialect identification results in average F1-score, precision and recall in % with 2 males and 2 females as the test set. The results are reported in terms of mean (μ) and standard deviation (σ) of three-fold cross-validation. The best performance is highlighted and represented in bold.

Features	F1-score ($\mu \pm \sigma$)	Precision ($\mu \pm \sigma$)	Recall ($\mu \pm \sigma$)
RMFCC	47.41±6.16	49.68±8.21	47.87±6.26
MFCC	53.79±3.91	56.57±3.16	55.28±3.20
SDC	65.15±3.59	65.99±3.61	65.05±3.92
MFCC+RMFCC	65.06±2.60	66.54±2.64	65.09±2.57
SDC+RMFCC	68.10±3.42	69.21±3.81	68.10±3.59
MFCC+SDC+RMFCC	**74.28±7.48**	**74.88±7.23**	**74.31±7.43**

Table 5.5 which follows the same trend as of Table 5.4. The classification result shows that the excitation source information, i.e., RMFCC feature, differentiates the three Ao dialects. Hence, it motivated us to explore the attention-based CNN-BiGRU

model to automatically learn the proposed time-frequency representations using PasL corpus with the motive of evaluating the proposed approach in a more generalized scenario.

5.7.2 Classification results using attention-based CNN-BiGRU

Table 5.6 shows the classification performance on speech segments of 1 second duration of the original data. Means (μ) and standard deviations (σ) of accuracy and F1-score, calculated from four-folds, are reported in the table. The proposed features S_{ilpr} and S_{LP-gm} gives a decent average F1-score of 43.68% and 44.73%, respectively. It is noticed that the performance of S_{ilpr} and S_{LP-gm} is better than the baseline feature MFCC. This indicates the importance of the proposed feature for the Ao DID task. It is also to be noted that the average performances of S_{ilpr} and S_{LP-gm} are higher than the chance probability, which encouraged us to explore the complementary information of the source features in combination with the VT features.

Table 5.6: Mean (μ) and standard deviation (σ) from four-fold cross-validation in classification of Ao dialects for 1 sec segment duration in the **original data**. ILPR-LMS, LP-gammatonegram, and LMS are represented by S_{ilpr}, S_{LP-gm}, and S_{lms}, respectively. The best performance is highlighted and represented in bold.

	Features	Accuracy ($\mu \pm \sigma$)	F1-score			
			Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
Baselines	MFCC	43.83±2.46	35.67±18.03	37.30±13.57	52.97±3.25	41.98±3.35
	S_{lms}	48.69±6.00	40.34±9.36	45.24±15.98	52.56±12.82	46.04±6.60
	S_{lms}+MFCC	49.80±5.77	40.36±10.58	45.81±15.40	56.15±10.04	47.44±6.19
Proposed	S_{ilpr}	46.01±3.83	29.10±18.19	53.43±5.36	48.50±5.56	43.68±4.45
	S_{LP-gm}	46.79±5.49	36.54±14.66	50.82±14.06	46.83±16.82	44.73±6.42
	S_{ilpr}+S_{LP-gm}	48.73±2.47	32.95±15.79	57.41±9.25	48.63±12.08	46.33±2.19
	S_{ilpr}+S_{lms}	49.94±3.53	35.91±13.11	53.56±9.35	52.77±10.65	47.41±2.82
	S_{LP-gm}+S_{lms}	49.59±5.89	39.82±11.22	46.91±16.80	53.16±13.98	46.63±6.58
	S_{ilpr}+S_{LP-gm}+S_{lms}	51.05±4.37	38.16±13.65	52.60±12.81	53.77±13.70	48.18±4.42
	S_{ilpr}+S_{LP-gm}+S_{lms}+MFCC	**52.52±4.71**	**40.12±12.26**	**51.93±13.10**	**59.03±9.13**	**50.36±4.75**

From Table 5.6, it is observed that when the source features are combined (S_{ilpr} +

S_{LP-gm}), it outperforms the standalone baseline features. Similarly, when the proposed feature (S_{LP-gm}) is combined with the VT feature, it gives an improved average F1-score. The best performance is obtained for the combination of $S_{ilpr} + S_{LP-gm} + S_{lms} + MFCC$ with an average F1-score of 50.36%. However, the performances are low and might be attributed to the following reasons. Firstly, speaker-independent dialect identification is more challenging in comparison to speaker-dependent. Secondly, the limited amount of speech data for training the classifier. Furthermore, it is well known that deep learning models do not generalize well on small datasets and that there may be overfitting issues. In order to compensate for the lack of data, speech data is augmented with two types of augmentation approaches.

5.7.3 Classification results after data augmentation

In this work, we argue that data augmentation specifically helps increase the accuracy of CNN-BiGRU based classification. In order to confirm that, we compared the effect of data augmentation on a traditional i-vector system and a CNN-BiGRU system with baseline MFCC features. Models were built with and without augmented data and were tested using the original speech data. The results as reported in Table 5.7 shows that when models are built with augmented data, average F1-scores for the i-vector and CNN-BiGRU systems are improved from 31.73% to 34.61% and 41.98% to 47.21%, respectively. Hence it is seen that data augmentation has improved the performance of the CNN-BiGRU system better than the traditional i-vector system.

After confirming the usefulness of augmented data, training is done using the original and the augmented speech data. The trained models were tested with the original speech data. Figure 5-9 shows the average F1-score plot for 1 sec segment duration with and without data augmentation. It is observed that data augmentation results in better performance for all the features, along with the proposed S_{LP-gm} feature. The highest average F1-score is achieved by the combination of source and

Table 5.7: Mean (μ) and standard deviation (σ) of Ao dialect classification performance for 1 sec segment duration from four-fold cross-validation using original and augmented data with i-vector and CNN-BiGRU classifiers for the baseline MFCC feature.

Feature	Training	Classifier	Accuracy ($\mu \pm \sigma$)	F1-score			
				Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
MFCC	Original data	i-vector	32.44±1.12	33.92±4.05	26.92±5.10	34.37±4.46	31.73±1.41
		CNN-BiGRU	43.83±2.46	35.67±18.03	37.30±13.57	52.97±3.25	41.98±3.35
	Original+	i-vector	36.01±1.09	37.67±3.66	35.63±2.88	30.54±6.38	34.61±1.18
	Augmented data	CNN-BiGRU	48.89±2.38	37.69±12.33	42.33±10.29	61.61±5.52	47.21±2.01

VT features, where data augmentation improved the identification rate from 50.36% to 57.43%. It was also noticed that data augmentation improved the performance of the source features more than the VT features. Without any combination, when VT features, MFCC, and S_{lms} were considered, data augmentation improved their performances by 12.65% and 9.78%, respectively. On the other hand, data augmentation improved the performance of the source features, S_{ilpr} and S_{LP-gm}, by 22.02% and 17.23%, respectively. It was also noticed that data augmentation reduced the σ values, possibly due to the increase in data size.

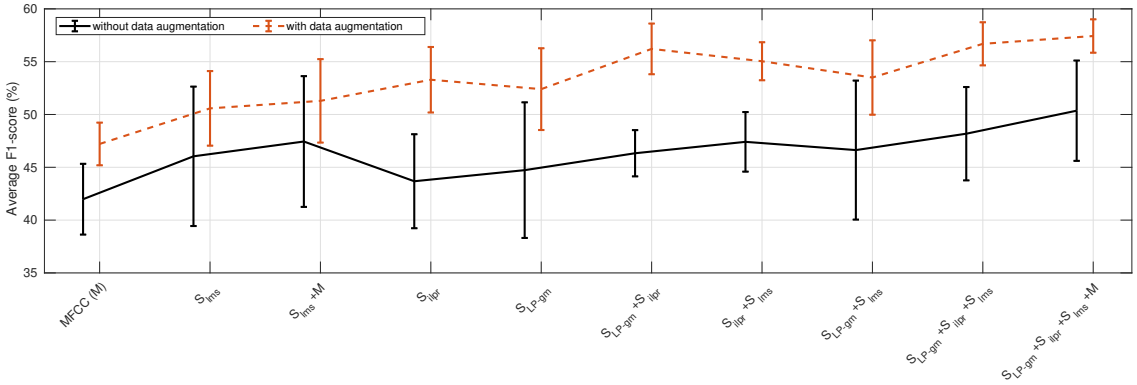


Figure 5-9: Classification performance of Ao dialects in 1 sec segment duration with and without data augmentation, indicated by orange and black lines, respectively.

The rest of the experiments in this chapter are conducted by incorporating the augmented data in the speech PasL database. All these results were computed using

the architecture described in sub-section 5.3.4. However, for the remaining experiments, hyper-parameter tuning is performed to obtain an optimized architecture.

5.7.4 Performance using optimized architecture

The experimental setup for evaluating optimized architecture is same as that discussed in sub-section 5.6.2. To obtain the classification results, a four-fold speaker-independent cross-validation training technique is used. The classification performance after parameter tuning for 1 sec segment duration is presented in Table 5.8. It is observed that the proposed features S_{ilpr} and S_{LP-gm} performs decently in discriminating the Ao dialects. The performance of S_{ilpr} and S_{LP-gm} in terms of average F1-score are also higher than MFCC and S_{lms}. This signifies the importance of the proposed features for the DID task. Also, as the individual features give a decent performance, feature combination is further explored. For two features combination, the features are fused at score level and are computed as shown in equation 5.8, where S^E_{f1} and S^E_{f2} are the prediction scores obtained for features $f1$ and $f2$.

$$S^E_1 = \alpha_1 S^E_{f1} + (1 - \alpha_1) S^E_{f2} \tag{5.8}$$

The value of α_1 varies from 0-1. For instance, the plot in Figure 5-10 shows the combination of source features, S_{ilpr} and S_{LP-gm}. The plot shows the average F1-score with different α_1 values where the highest performance is obtained for $\alpha_1 = 0.5$. It is also noticed when the source features are combined ($S_{ilpr} + S_{LP-gm}$), the performance is improved. Accordingly, for the three features combination, the prediction scores from equation 5.8 (S^E_1) and feature $f3$ (S^E_{f3}) are computed as shown in equation 5.9.

$$S^E_2 = \alpha_2 S^E_1 + (1 - \alpha_2) S^E_{f3} \tag{5.9}$$

In the same manner, for four feature combinations, the prediction scores from equation

5.9 (S_2^E) and feature $f4$ (S_{f4}^E) are computed as shown in equation 5.10.

$$S_{final}^E = \alpha_3 S_2^E + (1 - \alpha_3) S_{f4}^E \tag{5.10}$$

In equation 5.10, S_{final}^E provides the final predicted score for fusion of four features. It is observed that the highest performance is achieved when the excitation source features are combined with the VT features. There is an improvement of $\approx 3\%$ for the combination features i.e. $S_{ilpr} + S_{LP-gm} + S_{lms} + MFCC$ after tuning the parameter compared to the result shown in Figure 5-9.

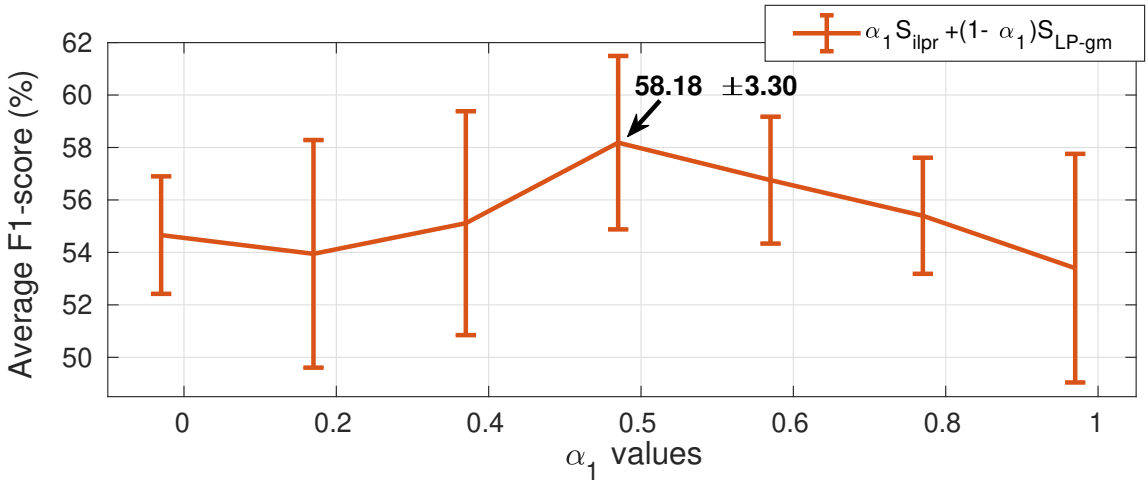


Figure 5-10: Performance of various linear combination weights (α_1) assigned to S_{ilpr} combined with S_{LP-gm}.

5.7.5 Classification results using various segment duration

Dialect perception by Ao speakers

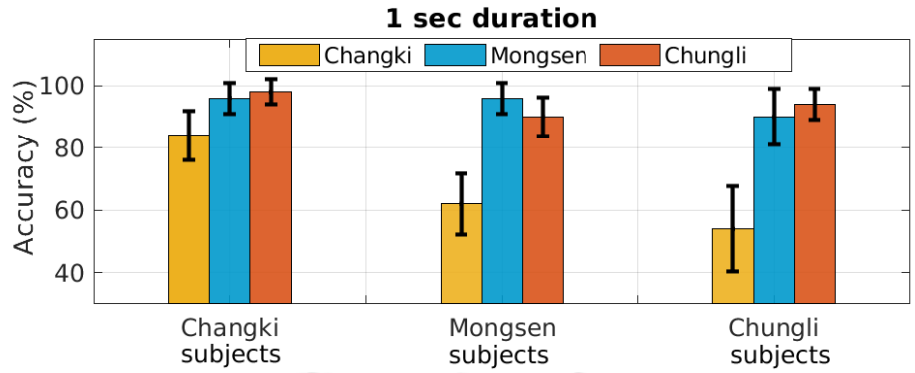
In order to see how utterance length affects dialect identification by the native speakers of Ao, we conducted a perception test as detailed in sub-section 5.6.4. As shown in Figure 5-11 (a), when the subjects were given a short segment of 1 sec, the Changki subjects performed the best, achieving an accuracy of 84%, 96%, and 98% in identifying Changki, Mongsen, and Chungli dialects, respectively. In identifying Changki,

Table 5.8: Mean (μ) and standard deviation (σ) from four-fold cross-validation in classification of Ao dialects for 1 sec segment duration **after hyper-parameter tuning**. ILPR-LMS, LP-gammatonegram, and LMS are represented by S_{ilpr}, S_{LP-gm}, and S_{lms}, respectively. The best performance is highlighted and represented in bold.

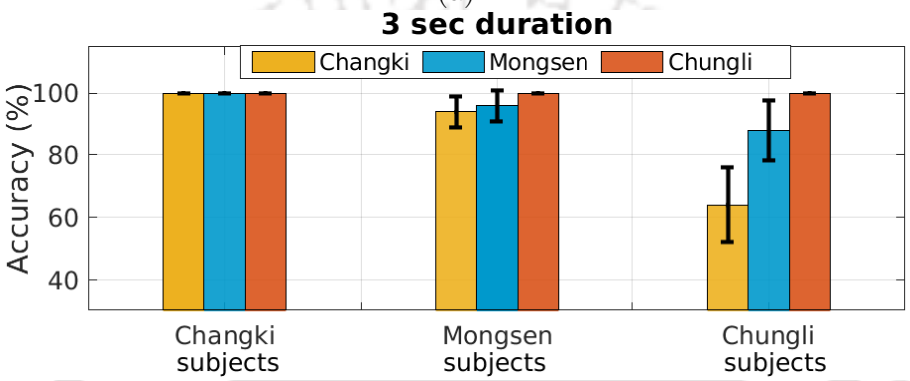
	Features	Accuracy ($\mu \pm \sigma$)	F1-score			
			Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
Baselines	MFCC	49.44±5.78	38.20±20.54	38.27±11.59	62.04±6.42	46.17±5.24
	S_{lms}	53.89±4.59	44.18±13.76	52.02±3.05	62.00±2.40	52.73±4.84
	S_{lms}+MFCC	54.84±4.25	44.47±14.10	51.58±3.70	64.02±2.35	53.36±4.33
Proposed	S_{ilpr}	54.84±2.44	53.01±5.12	57.78±5.05	53.20±5.24	54.66±2.24
	S_{LP-gm}	54.93±4.54	43.62±18.57	58.82±7.17	57.76±8.83	53.40±4.36
	S_{ilpr}+S_{LP-gm}	58.84±3.40	53.15±11.69	62.14±6.90	59.27±8.73	58.18±3.30
	S_{ilpr}+S_{lms}	56.67±1.79	53.48±6.01	58.75±4.06	57.56±4.39	56.60±1.76
	S_{LP-gm}+S_{lms}	55.90±4.73	43.95±18.66	59.38±6.16	59.77±7.99	54.37±4.48
	S_{ilpr}+S_{LP-gm}+S_{lms}	59.86±3.68	52.31±13.11	62.18±4.12	62.98±6.67	59.16±3.42
	S_{ilpr}+S_{LP-gm}+S_{lms}+MFCC	**60.27±3.11**	**51.85±12.59**	**60.31±4.19**	**65.83±5.77**	**59.33±2.44**

Mongsen, and Chungli dialects, the Mongsen subjects had an accuracy of 62%, 96%, and 90%, respectively; the Chungli subjects had an accuracy of 54%, 90% and 94%, respectively. Except for the Changki subjects, the other two groups' ability to identify the Changki dialect was lower. When the stimuli duration was 3 sec, the Changki subjects were able to identify all the three dialects with an accuracy of 100%, as shown in 5-11 (b). Similarly, Mongsen subjects also improved their dialect identification with the increased stimuli duration. However, with a longer duration, the Chungli subjects could identify their own dialect with an accuracy of 100%. Nevertheless, their identification of the other dialects was reduced.

It is seen that increased duration of stimuli helped in the identification of the dialects by Ao native speakers. However, in the perception test reported in this section, the pattern of identification of other dialects is not the same across the Ao speakers. We assume that such difference in dialect identification has to do with the familiarity of the subjects with the other dialects, further discussed in section 5.8.



(a)



(b)

Figure 5-11: Demonstrating the results of identifying Ao dialects based on the perceptual test. The perception test is performed on two different segment durations: (a) 1 sec and (b) 3 sec.

Automatic dialect classification

The classification results in terms of average F1-score modeled on different segment duration for excitation (Exc.) and VT features are shown in Figure 5-12. It is observed from the plot that the highest performance is achieved in 6 sec segment duration for MFCC, S_{lms} and S_{ilpr}. In contrast, the performance for S_{LP-gm} decreases after 3 sec segment duration. Also, an important point to observe is that as μ increases, σ also increases significantly for S_{lms}, S_{ilpr} and S_{LP-gm}. It is also noticed that the σ value is lowest in 1 sec segment duration. However, the next best performance is obtained in 3 sec with reduced σ values compared to the 6 sec segment duration. Thus, henceforth results of automatic DID in 3 sec segment duration are reported in this chapter, however, results for all durations are reported in chapter 7.

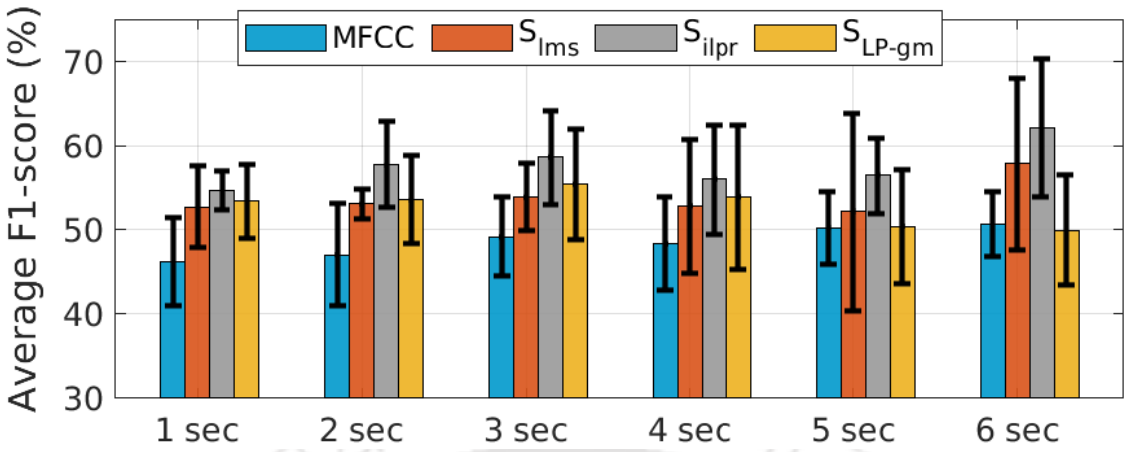


Figure 5-12: Classification performance for 1, 2, 3, 4, 5 and 6 sec segment duration. The best performance is in 6 sec. However, 3 sec is comparable as it has high μ with low σ.

Table 5.9 shows the performance of combined features in 3 sec segment duration. It is observed that the individual source features (S_{ilpr} and S_{LP-gm}) and their combination outperform the baseline VT features, MFCC, and S_{lms}. Therefore, we attempted combining the source and the VT features (S_{ilpr}+S_{LP-gm}+S_{lms}+MFCC), which yielded the highest performance score of 61.46%. This signifies the importance of excitation source information in identifying the three Ao dialects.

5.8 Discussion and summary

The results of this study confirm that the inclusion of source features improve DID in Ao. A DID task is performed in the three dialects of Ao using GMM classifier, where the fusion of MFCC, SDC and RMFCC yielded an accuracy of 74.31% from three-fold speaker-independent training strategy. Further, to assess the performance of Ao DID system in a generalised framework, attention-based CNN-BiGRU model is proposed. As shown in Table 5.9, inclusion of source features improved the accuracy in DID from 53.84% to 61.46%. In dialect identification, the dialects to be classified are usually very similar to each other. Hence, DID tasks usually report lower accuracy

Table 5.9: Classification performance of Ao dialects for 3 sec duration. Excitation features, vocal tract features, ILPR-LMS, LP-gammatonegram, and LMS are represented by Exc., VT, S_{ilpr}, S_{LP-gm}, and S_{lms}, respectively. The best performance is highlighted and represented in bold.

	Features	Accuracy ($\mu \pm \sigma$)	F1-score			
			Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
Exc.	S_{LP-gm}	57.44±5.09	47.74±21.20	59.64±16.77	58.83±11.50	55.40±6.60
	S_{ilpr}	59.45±5.33	50.53±13.07	64.44±8.09	60.89±11.79	58.62±5.60
	S_{ilpr}+S_{LP-gm}	60.66±5.60	51.14±15.94	65.59±9.65	62.13±12.50	59.62±6.24
VT	MFCC	51.89±4.48	38.62±22.84	43.06±12.37	65.76±9.93	49.14±4.69
	S_{lms}	55.27±3.19	44.80±18.37	54.23±7.60	62.51±5.05	53.84±4.02
	S_{lms}+MFCC	57.10±2.63	46.66±17.04	54.24±7.17	65.71±4.33	55.53±3.11
Exc.+VT	S_{LP-gm}+S_{lms}	58.45±4.98	47.40±20.67	60.09±14.46	62.57±11.01	56.69±6.07
	S_{ilpr}+S_{lms}	60.46±3.68	50.79±13.35	64.40±7.54	63.18±11.21	59.46±3.97
	S_{ilpr}+S_{LP-gm}+S_{lms}	61.87±4.69	51.85±16.27	66.01±9.03	64.20±12.38	60.69±5.31
	S_{ilpr}+S_{LP-gm}+S_{lms}+MFCC	**62.67±3.94**	**52.33±13.41**	**64.52±9.65**	**67.54±10.67**	**61.46±4.22**

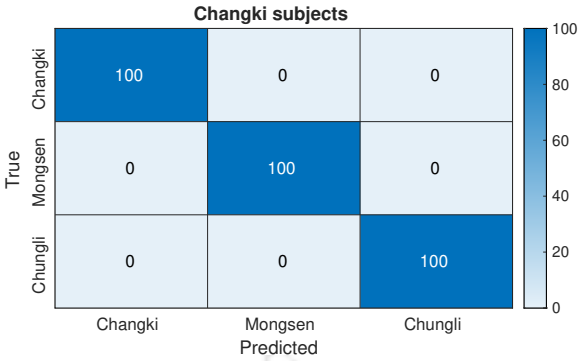
scores in the range of 60% to 70% [61, 62, 46, 42]. Dialect classification in this study also benefited from data augmentation as there was an improvement of 14% with the inclusion of augmented data in training.

In the current study, data augmentation improved the performance of the source features more than the VT features. The source features are noise robust compared to the VT features, as shown in *Sarma et al.* [101]. Hence, the noise induced augmented data in this study may have helped in preserving the dialect-specific prosodic information, improving dialect classification.

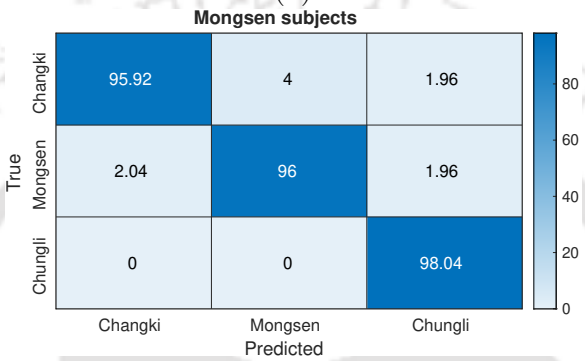
Apart from reporting an improved methodology for automatic dialect classification in Ao, we also have a few interesting observations. We observed that the sociolinguistic status of the three dialects has influenced the human perception and automatic identification of the dialects. In case of Ao, the Chungli dialect is considered to be the prestige variety of the language. Hence, the written form of the language uses the Chungli dialect. As a result of this, both Mongsen and Changki speakers are familiar with the Chungli dialect and are also capable of speaking it. On the other

hand, Chungli speakers can only speak their own dialect. Mongsen speakers are familiar with the Changki dialect, whereas, Chungli speakers are more familiar with the Mongsen dialect than the Changki dialect. This aspect is reflected in the results of the perception test reported in sub-section 5.7.5, where it is seen that the Chungli speakers are the worst in identifying dialects that are not their own, as seen in Figure 5-13 (c). As both Mongsen and Changki speakers are familiar with the Chungli dialect apart from each others', given longer speech segments, they are able to identify all three dialects with very high accuracy. It is important to acknowledge this sociolinguistic phenomenon as it may influence the reading of text by the Changki and Mongsen speakers.

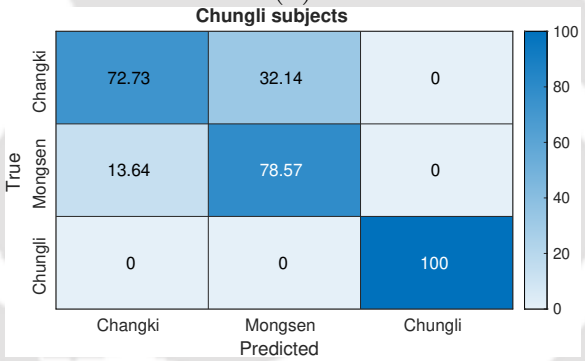
Figure 5-13 shows the confusion matrix of the perception tests conducted on the three dialect speakers along with the one obtained for automatic DID. Figure 5-13 (a)-(c) shows that for the Chungli and the Mongsen speakers, accuracy in identifying the three dialects follows a similar pattern, viz., Chungli>Mongsen>Changki. In case of automatic DID, using the S_{ilpr}+S_{LP-gm}+S_{lms}+MFCC features, a similar pattern is observed as in Figure 5-13 (d).



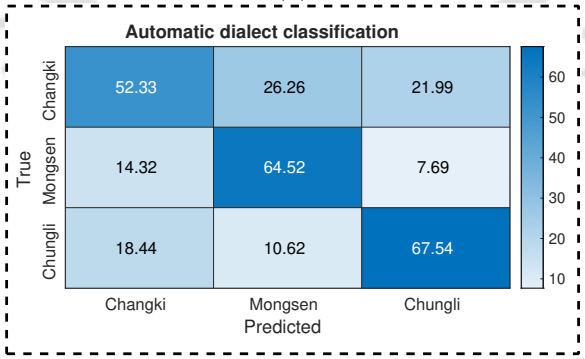
(a)



(b)



(c)



(d)

Figure 5-13: Confusion matrices for 3 sec segment duration in F1-score from the perception tests conducted on speakers of the three Ao dialects (a-c) and, from automatic classification of the three Ao dialects (d).

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 6

Prosodic Feature Based Dialect Identification

Overview

This chapter reports an automatic dialect identification system in the Ao language using prosodic features. Prosodic characteristics are believed to have an essential role in tonal languages. In this direction, the current work focuses on investigating the prosodic characteristics to build a discriminative system in identifying the three Ao dialects. The statistical and Low-Level Descriptors (LLD) of prosodic features are used in this chapter. The prosodic features such as F_0, loudness, shimmer, jitter, voiced and unvoiced segment length, etc., are utilized in this study. The experiments are conducted using SVM and attention-based Bi-GRU classifiers in trisyllabic words (TriW corpus) and passage level (PasL corpus) datasets, respectively. The statistical prosodic features are categorized into three sets: F_0Semitone, loudness, and Voice Quality and Temporal (VQT) features. The VQT feature set is the best performing prosodic feature in the SVM based classification, which captures the temporal information. Additionally, the combination of prosodic features outperforms the MFCC (baseline) feature. The statistical analysis also shows that the VQT features are statistically significant. Accordingly, in order to encapsulate the temporal characteristics, an attention-based Bi-GRU using LLD prosodic features is used for a DID task in Ao using the original speech data. The best performance is achieved for the combination of prosodic and vocal tract features. It is also noticed that the average $F1$-score of the combined features improved by 28.58% after augmenting the data. The performances of SVM and attention-based Bi-GRU classifiers indicate the significance of prosodic information in classifying the three Ao dialects.

6.1 Introduction

This chapter describes the design of an automatic DID system in the Ao language employing multiple prosodic feature parameters. Given that Ao is a tonal language,

prosodic traits that capture change in various tonal aspects are thought to be essential in distinguishing the varieties of Ao language. As a result, this chapter explores the prosodic features to automatically identify the three dialects of Ao.

As per the literature, majority of the works have explored spectral features such as MFCC and SDC. A number of works have studied the prosodic feature viz., F_0, energy, intensity, and duration. However, a detailed study of prosodic characteristics has not been presented in the literature for dialect identification. As Ao is a tonal language, it is believed that prosodic characteristics will play a vital role in discriminating the dialects. The variations in the prosodic information may capture the potential differences across the three dialects of Ao. Hence, to see the effectiveness of prosodic information in Ao, multiple prosodic features are extracted that have not been explored yet in DID tasks [67]. Also, most of the works studied previously are based on high-resource languages. In case of the Arabic language, the Arabic dialects such as Modern Standard Arabic (MSA), Iraqi Arabic, and Levantine Arabic are standard dialects of different countries. These dialects are used in broadcast news with available written scripts. However, Ao is a low-resource language where the Chungli dialect is known to be the standard dialect of the language. Hence, the Chungli dialect is used in all formal occasions and gatherings. In particular, all written text is available only in the standard dialect. As such, speech analysis and modeling becomes more challenging in Ao. Accordingly, an automatic Ao DID system is attempted using prosodic information in this work. The contributions of this chapter are listed below.

- The present work proposes the use of prosodic information for DID in Ao. First, the statistical prosodic feature is utilized with Support Vector Machine (SVM) based classifier. The classification results establish the importance of prosodic characteristics for the current task.

- Statistical significance analysis is also conducted to further analyze the efficacy

of statistical prosodic features. The result indicates that the VQT based prosodic features are more significant in the current DID task. This motivates to explore temporal information of different prosodic features for the task.

- The temporal information of prosodic features is learned using an attention-based Bi-GRU model. The Low-Level Descriptors (LLD) of prosodic information is used with an attention-based Bi-GRU classifier to classify the three Ao dialects. The attention mechanism is performed along the feature dimension. The attention provides higher weights to the features that carry prominent discriminative information across the dialects.

The remaining paper is organized as follows. Section 6.2 gives a brief description of the speech corpus for this work. The proposed approach is described in section 6.3. Section 6.4 presents the baseline methods. Section 6.5 and section 6.6 discusses the experimental setup and results. Finally, the work is summarized in section 6.7.

6.2 Speech corpus

This chapter utilizes two speech datasets from the CMC-Ao database, as described in chapter 3. The TriW corpus is used to employ DID tasks in Ao using SVM. While, the PasL corpus utilizes attention-based Bi-GRU model.

6.3 Proposed work of Ao DID system

The proposed framework of the Ao DID system is illustrated in Figure 6-1. The input speech signal is initially pre-processed by resampling to 16 kHz, applying z-score normalization and detecting the speech region by removing the silence regions. Next, the pre-processed speech is passed through the prosodic feature extraction block. The

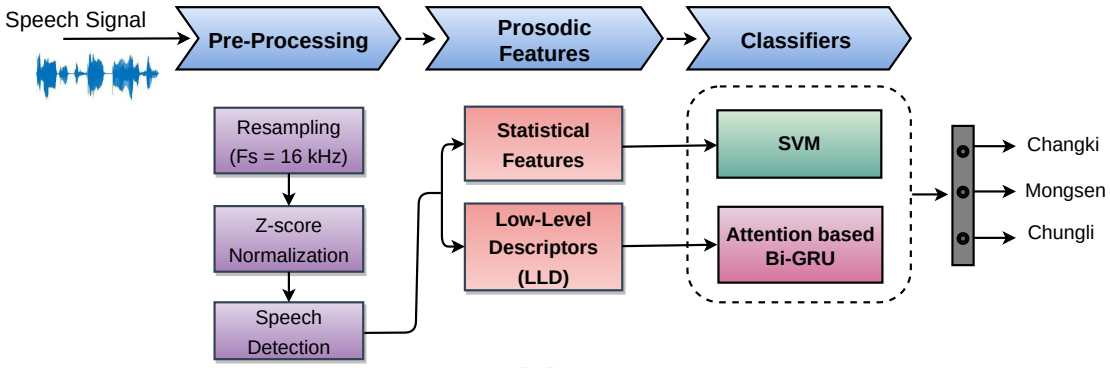


Figure 6-1: Overall framework of Ao DID system

extracted features are then fed to the classifiers for the three-class classification task. The description of prosodic features and classifiers are presented next.

6.3.1 Prosodic features

The variations in suprasegmental characteristics of the speech signal are believed to carry potential differences across the three dialects. As Ao is a tonal language, the suprasegmental features capturing variation in different aspects of tone are expected to be helpful for the task. Therefore, different statistical features of F_0 semitone along with jitter and shimmer are explored. Similarly, the present work also focuses on exploring other prosodic features, such as loudness, voice quality and temporal features across the dialects. In this chapter, the prosodic features are extracted using the openSMILE toolkit [67, 102, 103]. The prosodic features are extracted on two levels: Functionals and LLD. In the case of functionals [102], 30 statistical prosodic features in terms of mean (μ) and its standard deviation (σ) are extracted from TriW corpus and are categorized into 3 groups:

(i) **F_0Semitone (F_0ST) features**: 10 statistical features, namely, percentile 20, 50, and 80 of F_0ST, μ and σ, rising slope and falling slope of F_0ST.

(ii) **Loudness features**: 10 statistical features viz., percentile 20, 50, and 80 of loudness, μ and σ, rising slope and falling slope of loudness.

(iii) **Voice Quality and Temporal (VQT) features**: Voice quality comprises of 4 statistical features viz., μ and σ of jitter, shimmer. While, the temporal features include 6 statistical features, namely, loudness peaks per sec, voiced segments per sec, μ and σ of voiced segment length and unvoiced segment.

In the case of LLD [102, 103, 104], 14 prosodic features are extracted from PasL corpus and are categorized into 2 groups:

(i) Prosodic set 1 (P1):

- **Loudness**: a measure of perceived signal intensity from an auditory spectrum.
- F_0**ST**: the logarithmic F_0 on a semitone frequency scale, beginning at 27.5 Hz.
- **Jitter**: deviations in each successive F_0 period length.
- **Shimmer**: difference between the peak amplitudes of successive F_0 phases.
- **Harmonic-to-Noise Ratio (HNR)**: relation between the energy in harmonic components and the energy in noise-like components.
- **Harmonic difference H1-H2**: the energy-to-energy ratio of the first and second F_0 harmonics (H1 and H2, respectively).
- **Harmonic difference H1-A3**: the energy-to-energy ratio of the first harmonic (H1) and the highest harmonic in the third formant range (A3).

(ii) Prosodic set 2 (P2):

- F_0**final**: the smoothed F_0 contour.
- **Sum of RASTA-filtered auditory spectrum**: RASTA-filtered loudness.

- **Root Mean Square (RMS) energy**: the energy of a signal that corresponds to its total magnitude.

- **Zero-Crossing Rate (ZCR)**: the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.

- **Probability of voicing**: the voicing probability of the final F_0.

- **Log HNR**: logarithmic HNR.

- **Jitter (δ)**: the differential frame-to-frame jitter.

6.3.2 Attention-based Bi-GRU

The architecture of the attention-based Bi-GRU model is illustrated in Figure 6-2. This architecture is motivated from the previous chapter (Chapter 5). Temporal information plays a vital role in capturing the prosodic details of a speech signal. Therefore, the proposed work is encouraged to use Bi-GRU based architecture to learn the temporal variations of different dialects. The proposed model uses an attention mechanism [93] along the feature direction. The attention mechanism gives higher weights to those features that provide essential information for the classification task. The model consists of two Bi-GRU layers with 128 units each. The output of the attention module and Bi-GRU is concatenated. The concatenated output is fed to two dense layers with 32 neurons each. All the dense layers have $ReLU$ activation and a dropout rate of 0.4. Finally, the output layer (size = 3) is activated with *Softmax* function. The model is trained for 50 epochs with a mini-batch size of 33. An early stopping criteria is used to avoid overfitting of the model. The model is trained with categorical cross-entropy loss and an initial learning rate of 0.0001.

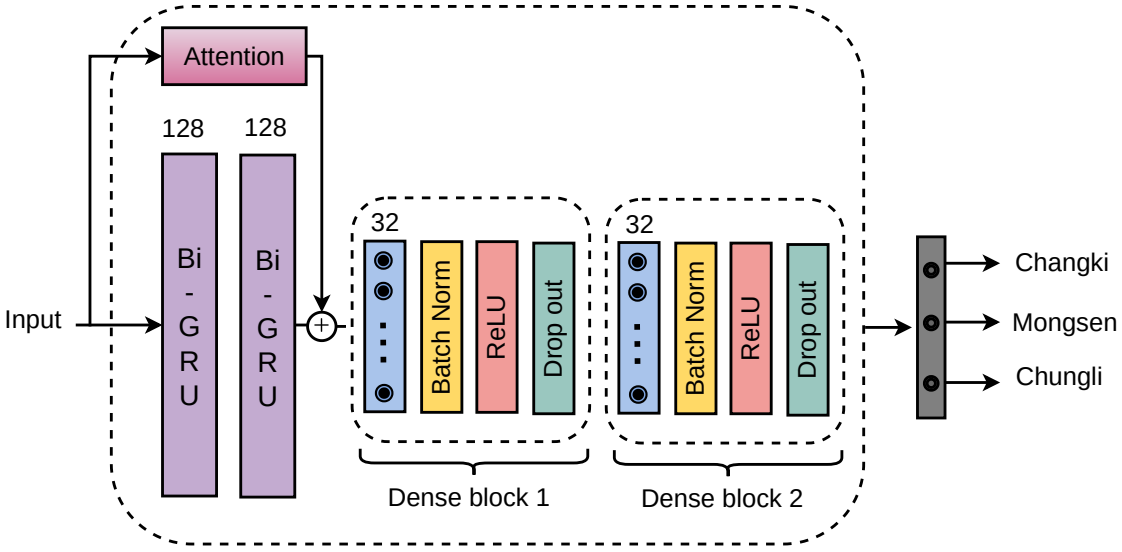


Figure 6-2: Architecture of attention-based Bi-GRU model

6.4 Baselines

MFCC feature is broadly used in DID tasks to capture the vocal tract information [25, 37, 35, 79]. Therefore, MFCC is considered as the baseline feature, which is extracted using the openSMILE toolkit [67]. The statistical features (μ and σ) are extracted from the trisyllabic words (TriW corpus). While, the 13-dimensional MFCC features with their Δ and $\Delta\Delta$ are extracted from the passage level data (PasL corpus) considered as MFCC-LLD.

SVM classifier with Radial Basis Function (RBF) kernel is trained using statistical prosodic features extracted from trisyllabic words (TriW corpus). The optimum values of the kernel parameters, c and γ are obtained using the grid-search mechanism. The parameters c and γ are considered in the range of $c = [10^{-1}, 10^{0}, \ldots 10^{+2}]$ and $\gamma = [10^{-3}, 10^{-2}, \ldots 10^{0}]$ for the grid search.

6.5 Experimental setup

This section describes the experimental setups for this chapter. The LLD features are calculated with a window size of 20 ms and a hop size of 10 ms.

6.5.1 SVM based classification

Initially, DID task in Ao is conducted with statistical prosodic features extracted from the trisyllabic words (TriW corpus). The trisyllabic data is divided into three non-overlapping folds consisting of 2 females and 2 males in each fold. Each fold consists of different sets of speakers resulting in a speaker-independent framework. The speech data from two folds are used for training, while the remaining fold is used for testing at every iteration. To demonstrate the usefulness of prosodic information in Ao, statistical features of F_0ST, loudness and VQT are used. The optimum values of c and γ parameters obtained after grid-search are 10 and 0.1, respectively.

6.5.2 Variable importance using SVM

In order to speed up training and result in better classification results, reducing the number of features plays an important role. Variable importance aids in identifying the main features for classification by excluding the features that are not relevant. Variable importance represents the statistical significance of each variable in the data in terms of its effect on the generated model. Figure 6-3 shows the variable importance for F_0ST, Loudness, VQT based prosodic features conducted using SVM with linear kernel. The variables are presented in order of descending importance for the proposed features.

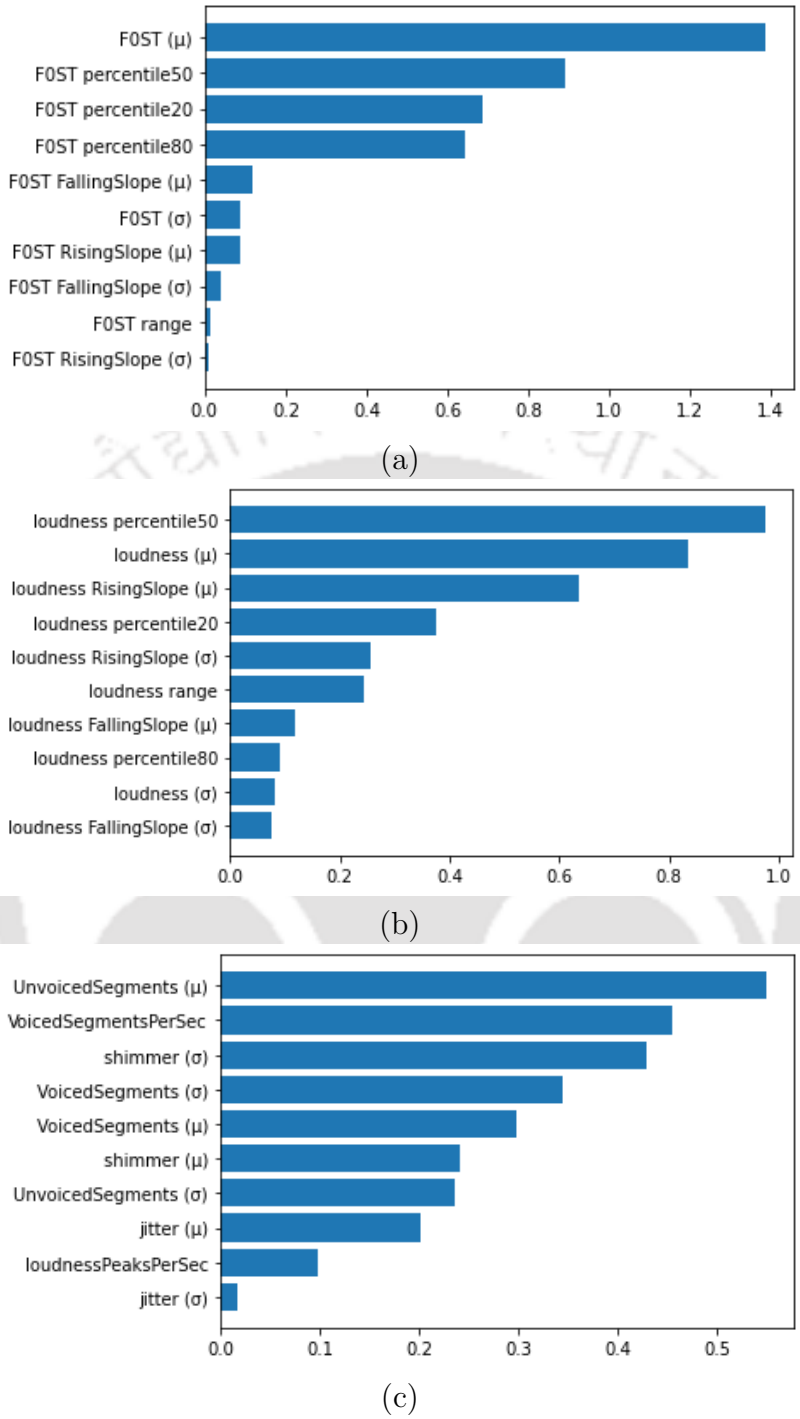


Figure 6-3: Feature importance plot using SVM (a) F_0ST, (b) Loudness, (c) VQT

6.5.3 Attention-based Bi-GRU Classification

The attention-based Bi-GRU classifier is trained separately for LLD prosodic features (P1 and P2), and 39-dimensional MFCC features. PasL corpus is used to train

the attention-based Bi-GRU model. A four-fold speaker-independent cross-validation training approach is conducted to obtain the classification results. Each fold consists of 1 female and 1 male. Further, the training set is split into a 70 : 30 ratio to get the train and validation set, respectively. The model is trained for 3 sec segment duration to utilize the temporal information of the speech data.

6.5.4 Data augmentation

To increase the dataset and avoid overfitting for the classification process, PasL corpus was augmented to telephonic and reverberated speech. G.191 software was implemented to convert the original speech data into telephonic speech [97]. A pipeline process reported in [98] is used for simulation. On the other hand, the dataset was also augmented to two types of reverberated speech using Roomsim toolbox [100]. The two categories of reverberated speech vary in terms of source and room sensor configurations. After data augmentation, the speech data resulted in $\approx$ 24 hours consisting of 384 passages across the three Ao dialects. A detailed description is presented in sub-section 5.6.3.

6.6 Results

This section reports the results of Ao DID to see the effect of variable importance using SVM. The section is followed by the classification results after variable importance with statistical analysis. This section also discusses the classification results obtained using attention-based Bi-GRU in original speech data and results after data augmentation.

6.6.1 SVM based classification results for the effect of variable importance

As explained in sub-section 6.3.1, SVM based classification is computed on the 3 feature groups: F_0ST, Loudness, and VQT based prosodic features (10 statistical features each). Figure 6-3 shows the variable importance of these features, where the SVM classification is performed for the top 4 features in each group. Table 6.1 shows the classification performance comprising of 10 and 4 feature sets. The results are presented in terms of μ and σ calculated from three-fold cross-validation performances. The table shows that F_0ST and VQT perform better with the 10 feature set. On the contrary, using the top 4 feature set, loudness results in a higher average F1-score. As a result, the remaining experiments employ the 10 statistical features for F0ST and VQT. For loudness, the top 4 statistical features are used.

Table 6.1: Classification performance of Ao dialects using statistical prosodic features in trisyllabic words (TriW corpus). The results are reported in terms of mean (μ) and standard deviation (σ) of three-fold cross-validation.

			F1-score			
	Features	Accuracy ($\mu \pm \sigma$)	Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
10 feats	F_0ST	**36.91±4.83**	**37.66±7.34**	**32.93±2.07**	**39.07±5.08**	**36.86±5.64**
	Loudness	38.84±3.64	39.87±5.90	30.61±6.44	44.98±1.61	36.96±4.16
	VQT	**46.08±1.65**	**41.30±1.93**	**43.88±1.86**	**52.41±1.84**	**44.16±1.35**
Top 4	F_0ST	30.15±4.93	28.82±10.13	24.15±7.33	32.36±6.46	28.44±4.02
	Loudness	**39.10±5.69**	**40.79±7.78**	**26.66±6.40**	**47.19±3.50**	**38.21±5.19**
	VQT	43.92±1.47	40.48±1.83	41.07±0.15	49.60±3.15	43.72±1.49

6.6.2 SVM based classification results after variable importance

The best performances for F_0ST, loudness, and VQT reported in Table 6.1 are further used for classification. Table 6.2 shows comparable average F1-scores obtained for F_0ST and loudness features. The MFCC feature outperforms the F_0ST and loudness features. However, the best performance is obtained for the VQT feature compared to other individual features. The decent performance of statistical prosodic features encourages to further explore them in combination.

Table 6.2: Classification performance of Ao dialects after variable importance. The results are reported in terms of mean (μ) and standard deviation (σ) of three-fold cross-validation. The best performance is highlighted and represented in bold.

Features	Accuracy ($\mu \pm \sigma$)	F1-score			
		Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
MFCC (M)	43.21±5.07	30.67±9.05	49.36±2.88	43.24±6.32	41.09±6.07
Loudness	39.10±5.69	40.79±7.78	26.66±6.40	47.19±3.50	38.21±5.19
F_0ST	36.91±4.83	37.66±7.34	32.93±2.07	39.07±5.08	36.86±5.64
VQT	**46.08±1.65**	**41.30±1.93**	**43.88±1.86**	**52.41±1.84**	**44.16±1.35**
F_0ST+Loudness	46.90±5.12	44.72±6.94	40.73±4.86	54.52±3.82	46.66±5.09
VQT+F_0ST	46.64±2.18	42.48±2.53	44.67±2.13	52.25±2.93	46.47±2.22
VQT+Loudness	46.94±2.52	42.88±2.90	44.01±2.33	53.36±2.62	46.75±2.57
VQT+Loudness+F_0ST	47.20±4.00	43.82±5.40	42.59±3.41	54.45±3.47	46.96±4.03
VQT+Loudness+F_0ST+M	**49.12±5.72**	**43.95±7.54**	**44.83±5.45**	**57.72±4.49**	**48.83±5.79**

For two features combination, the features are fused at score level and are computed as shown in equation 6.1.

$$S_1^{Ptri} = \alpha_{P_1^{tri}} S_{f1}^{Ptri} + (1 - \alpha_{P_1^{tri}}) S_{f2}^{Ptri} \tag{6.1}$$

where, S_{f1}^{Ptri} and S_{f2}^{Ptri} are the prediction scores obtained for features $f1$ and $f2$. The value of $\alpha_{P_1^{tri}}$ varies from 0-1. Comparable average F1-score for the combination of F_0ST+loudness, VQT+F_0ST, and VQT+loudness features are obtained and are higher than the individual MFCC feature. However, a lower σ is obtained for the VQT

feature compared to VQT+F_0ST and VQT+loudness. Accordingly, three features combination is computed as shown in equation 6.2.

$$S_2^{P^{tri}} = \alpha_{P_2^{tri}} S_1^{P^{tri}} + (1 - \alpha_{P_2^{tri}}) S_{f3}^{P^{tri}} \tag{6.2}$$

where, $S_1^{P^{tri}}$ is the score obtained from equation 6.1 and $S_{f3}^{P^{tri}}$ is the prediction score of feature $f3$. The best performance is obtained for the combination of VQT+loudness+F_0ST. An improvement of $\approx 6\%$ in average F1-score is observed for the VQT+loudness+F_0ST combination than individual VQT. This improvement is obtained for $\alpha_{P_2^{tri}} = 0.9$ value for the VQT feature. This implies that a higher weight is assigned to the VQT feature compared to other features. This signifies that VQT features are the most important features for the task. Moreover, higher performance for most statistical prosodic features than MFCC justifies the significance of the prosodic information for the current task.

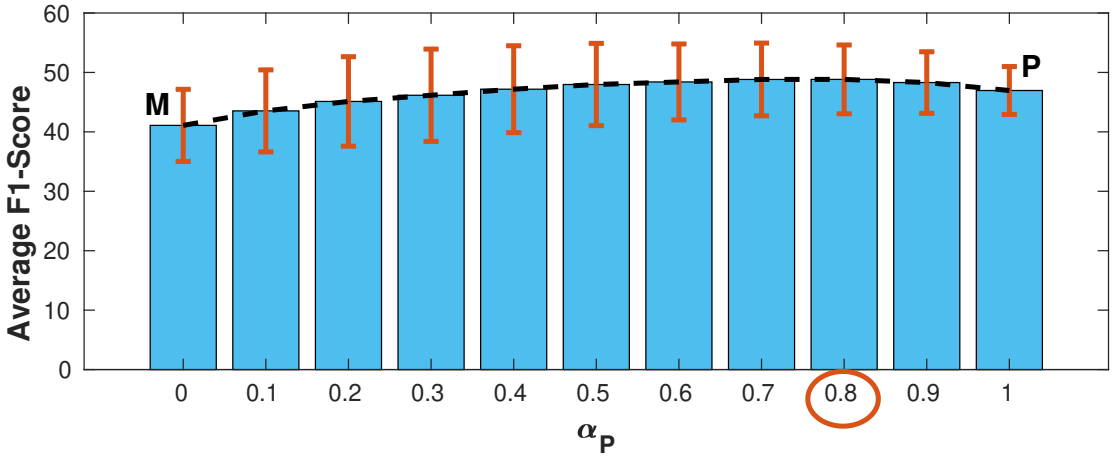


Figure 6-4: alpha variation for the 4 features combination reported in Table 6.2. M = MFCC, P = Prosodic.

In a similar way, the prosodic features (VQT+loudness+F_0ST) are finally fused with the MFCC feature to utilize the complementary information for the task. The

four feature combinations is obtained from equation 6.3.

$$S_{final}^{Ptri} = \alpha_P S_2^{Ptri} + (1 - \alpha_P) S_{f4}^{Ptri} \tag{6.3}$$

In equation 6.3, S_2^{Ptri} is the score obtained from equation 6.2, S_{f4}^{Ptri} is the prediction score of feature $f4$, and S_{final}^{Ptri} provides the final predicted score. The performance for this combination at different α_P is illustrated in Figure 6-4. The performance is better as the value of α_P increases, i.e., it gives higher weight to the prosodic features in comparison to the MFCC features. The best performance is obtained for $\alpha_P = 0.8$ (circled in red) as shown in Figure 6-4. Here, a weight of 0.8 is assigned to the VQT+loudness+F_0ST combination and 0.2 weight is assigned to MFCC. The performance obtained for $\alpha_P = 0.8$ is reported in Table 6.2. The combination of prosodic and MFCC features resulted in a further improvement of $\approx 3\%$ in the average F1-score in comparison to the prosodic (VQT+loudness+F_0ST) feature. These observations justify the effectiveness of prosodic features, specifically VQT, in identifying the three Ao dialects.

6.6.3 Statistical analysis

Analysis of variance (ANOVA) is conducted to observe the statistical significance of F_0ST, loudness, and VQT features. Table 6.3 shows the top 4 statistical significant F_0ST, loudness and VQT features with its p-value and F-value. These results validate the performances obtained from the SVM based variable importance experiment detailed in sub-section 6.5.2. The p-values indicate that the features are statistically different. Also, the F-value is related to the p-value inversely. Higher F-value indicates a significant p-value. It is noticed from the table that loudness and temporal features have the highest F-value. Hence, this motivated us to use Bi-GRU to learn the temporal context of the LLD prosodic features.

Table 6.3: Statistical significance analysis using ANOVA for the best four F_0ST, loudness, and VQT features. Results are reported in terms of F-value and p-value. The degree of freedom is represented by df.

Features	df	F-value	p-value
F_0ST (μ)	2	79.5	<0.001
F_0ST percentile 20	2	81.8	<0.001
F_0ST percentile 50	2	85.1	<0.001
F_0ST percentile 80	2	76.4	<0.001
Loudness (μ)	2	257.7	<0.001
Loudness percentile 20	2	157.4	<0.001
Loudness percentile 50	2	208.9	<0.001
Loudness rising slope (μ)	2	307.0	<0.001
shimmer (σ)	2	37.6	<0.001
voiced segments per sec	2	121.7	<0.001
voiced segments (σ)	2	36.6	<0.001
unvoiced segments (μ)	2	140.1	<0.001

6.6.4 Attention-based Bi-GRU classification results in the original speech

The model is trained for a segment duration of 3 sec to capture the temporal variations of the speech signal. Table 6.4 shows the classification performance using LLD features in the original speech data. From Table 6.4, it is observed that the prosodic features, specifically P2 give an improved performance in comparison to the results reported in Table 6.2. It is also noticed that the prosodic features outperform the baseline MFCC-LLD features. Additionally, for two features combination (P1+P2), the features are combined as given in equation 6.4.

$$S_1^P = \alpha_{P_1} S_{f1}^P + (1 - \alpha_{P_1}) S_{f2}^P \tag{6.4}$$

where, S_{f1}^{P} and S_{f2}^{P} are the prediction scores obtained from P1 and P2. Similarly, three features combination (P1+P2+MFCC-LLD) is computed as shown in equation 6.5.

$$S_{final}^{P} = \alpha_{P_final} S_{1}^{P} + (1 - \alpha_{P_final}) S_{f3}^{P} \tag{6.5}$$

where, S_{1}^{P} is the score obtained from equation 6.4 and S_{f3}^{P} is the prediction score for MFCC-LLD. The best performance is achieved for $\alpha_{P_{final}} = 0.98$, assigning higher weight to prosodic features. However, the performances are low, hence, speech data is augmented with two types of augmentation methods to make up for the limited speech data.

Table 6.4: Classification performance of Ao dialects using LLD features in 3 sec segment duration in the **original speech data** (PasL corpus). The results are reported in terms of mean (μ) and standard deviation (σ) of four-fold cross-validation. Prosodic set 1, prosodic set 2 are represented by P1 and P2, respectively. The best performance is highlighted and represented in bold.

Features	Accuracy ($\mu \pm \sigma$)	F1-score			
		Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
MFCC-LLD (M-LLD)	41.30±9.18	22.91±16.26	35.29±16.40	55.25±15.39	37.82±8.55
Prosodic set 1 (P1)	46.22±3.83	39.71±16.37	40.64±21.14	50.29±9.41	43.55±4.35
Prosodic set 2 (P2)	50.86±9.53	44.97±3.87	50.75±28.16	51.19±13.69	48.97±12.07
P1+P2	51.74±8.70	45.37±4.06	51.33±28.96	52.75±11.89	49.82±11.43
P1+P2+M-LLD	**52.00±9.01**	**45.49±3.55**	**51.82±29.30**	**53.10±11.84**	**50.13±11.75**

6.6.5 Attention-based Bi-GRU classification results after data augmentation

The classification performance using LLD features after data augmentation is shown in Table 6.5. For the classification process, training is done using the the original and augmented speech data. While the trained models are tested using the original speech data. The results reported in this chapter are for a segment duration of 3 sec, however,

the results of automatic DID for 1 - 6 sec segment durations are reported in chapter 7. From Table 6.5, it is observed that data augmented performances are better for all the features, along with the proposed prosodic features (P1 and P2) in comparison to the results reported in Table 6.4. The highest average F1-score is achieved for the combination of prosodic and MFCC-LLD features with $\alpha_{P_{final}} = 0.8$ using equation 6.5, assigning higher weight to prosodic features. Also, data augmentation improved the performance from 50.13% to 64.46%. It is also noticed that data augmentation improved the performance of the prosodic features, P1 and P2, by 19.86% and 30.22%, respectively. Hence, these results substantiate the importance of prosodic information by capturing dialect-specific characteristics to classify the three dialects of Ao.

Table 6.5: Classification performance of Ao dialects using LLD features in 3 sec segment duration **after data augmentation**. The results are reported in terms of mean (μ) and standard deviation (σ) of four-fold cross-validation. Prosodic set 1, prosodic set 2 are represented by P1 and P2, respectively. The best performance is highlighted and represented in bold.

Features	Accuracy ($\mu \pm \sigma$)	F1-score			
		Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
MFCC-LLD (M-LLD)	45.78±4.17	27.26±16.85	43.33±10.28	61.06±8.52	43.88±5.05
Prosodic set 1 (P1)	53.39±6.48	44.84±12.69	60.60±14.25	51.18±15.96	52.20±8.14
Prosodic set 2 (P2)	65.77±4.50	61.68±6.28	61.24±21.08	68.40±11.69	63.77±5.39
P1+P2	65.64±5.17	60.55±5.03	64.60±19.57	67.87±11.52	64.34±6.23
P1+P2+M-LLD	**65.69±5.31**	**60.10±5.41**	**64.53±19.46**	**68.76±11.01**	**64.46±6.31**

6.7 Discussion and summary

This chapter presents DID system to automatically identify the three Ao dialects using prosodic information. First, the statistical prosodic features extracted from trisyllabic words (TriW corpus) are used with SVM classifiers. Secondly, the attention-based Bi-GRU model learns the temporal context of LLD prosodic features. In addition,

statistical analysis is conducted for F_0ST, loudness and VQT features, confirming the importance of temporal features. The overall classification performance with SVM and attention-based Bi-GRU justify the efficacy of prosodic information for the current task. The important point to notice is that the Chungli dialect gives the highest performance compared to Changki and Mongsen dialects (Table 6.2, Table 6.4 and 6.5). The reason can be attributed to the following point. As the standard variety of the language is Chungli, the Changki and Mongsen speakers switch to the standard dialect in formal occasions. In addition, the written form is only available in Chungli, and the speech data collected was in read speech. This may have influenced the pronunciation of some Changki and Mongsen dialect speakers towards the standard dialect as described in section 4.6. The next chapter discusses the utilization of the proposed tonal, excitation, and prosodic features to build a combined system for Ao DID tasks.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 7

Combined Framework for Dialect Identification in Ao

Overview

The objective of this chapter is to develop a combine Ao DID system using the various approaches proposed in the preceding chapters. The combined system is made up of three separate systems: system 1 *(S*1*), system* 2 *(S*2*), and system* 3 *(S*3*). These three systems are fused at score level for the three-class classification task. The PasL corpus is utilized for all three systems. MFCC and LMS are considered as the baseline features. The experiments are conducted at various durations to determine the effect of segment duration. In the latter part of this chapter, a specific use case is examined to assess the practicability of the proposed approach.*

7.1 Introduction

This chapter presents a combined approach of an automatic Ao DID system with the approaches proposed in chapter 4, chapter 5, and chapter 6 using the PasL corpus (original and augmented speech data). Tonal features (derived from F_0) for the Ao DID system is proposed in chapter 4. The significance of source information in recognizing the three Ao dialects is reported in chapter 5. Prosodic features are explored in chapter 6 to confirm dialect-specific traits in the Ao DID task. These approaches exploit different characteristics of speech signal. The success of individual method encouraged the development of a combined DID system.

7.2 Combined Ao DID system

The combined Ao DID system comprises of the systems proposed in chapter 4, chapter 5, and chapter 6. In chapter 4, F_0 features were restricted to TriW corpus. However, in this chapter, F_0 is extracted along with its Δ and $\Delta\Delta$ derivatives using the PasL corpus. System 1 (S1) is composed of 3-dimensional features (F_0+ΔF_0+$\Delta\Delta F_0$) modeled with the classifier from chapter 6. System 2 (S2) comprises of the model

proposed in chapter 5 with ILPR-LMS and LP-gammatonegram features. System 3 (S3) is composed of the model proposed in chapter 6 with prosodic set 1 (P1) and prosodic set 2 (P2). The overall Ao DID framework for identifying the three Ao dialects is depicted in Figure 7-1. The initial pre-processing steps involve resampling the input speech signal to 16 kHz, implementing z-score normalization, and identifying the speech region by removing the silence regions. Following that, the features are extracted and fed to the classifier. Attention-based Bi-GRU classifier is used to model F_0, P1 and P2 features. While LP-gammatonegram and ILPR-LMS are input into the attention-based CNN-BiGRU classifier. Score fusion is computed from the prediction scores generated from the three systems (S1, S2 and S3) for three-class classification task.

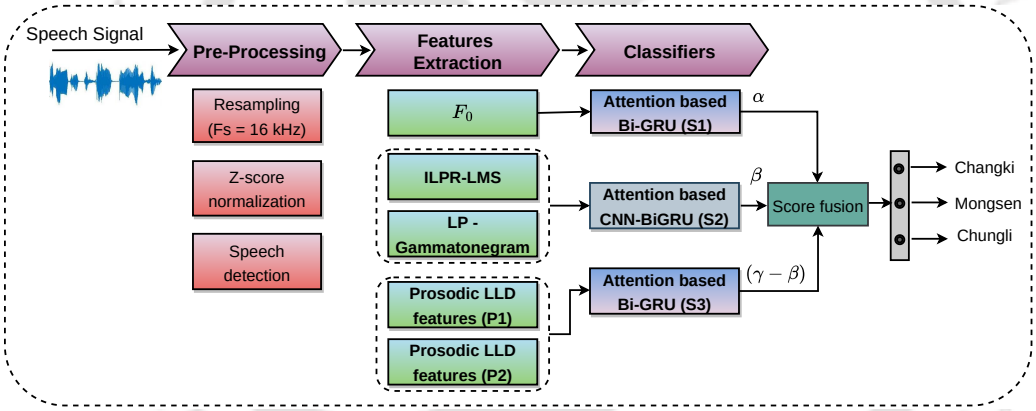


Figure 7-1: Combined framework of Ao DID system

7.3 Experiments

MFCC and LMS are considered as the baseline features with attention-based CNN-BiGRU as the classifier. This experiment is conducted using the PasL corpus (original and augmented speech data), where the speech data is divided into four speaker-independent non-overlapping folds for all the three systems, S1, S2 and S3. The fold is divided so that each fold contains 1 male and 1 female speakers. Each iteration

Table 7.1: Classification performance of Ao dialects at various segment duration in average F1-score. The results are reported in terms of mean (μ) and standard deviation (σ) for four-fold cross-validation. Vocal Tract, LMS, ILPR-LMS, LP-gammatonegram, Prosodic set 1, Prosodic set 2, System 1, System 2, System 3 are represented by VT, S_{lms}, S_{ilpr}, S_{LP-gm}, P1, P2, S1, S2 and S3, respectively.

	Features	Average F1-score					
		1 sec ($\mu \pm \sigma$)	2 sec ($\mu \pm \sigma$)	3 sec ($\mu \pm \sigma$)	4 sec ($\mu \pm \sigma$)	5 sec ($\mu \pm \sigma$)	6 sec ($\mu \pm \sigma$)
VT	MFCC (M)	46.17±5.24	47.04±5.99	49.14±4.69	48.33±5.50	50.17±4.25	50.68±3.84
	S_{lms}	52.73±4.84	53.04±1.75	53.84±4.02	52.86±7.99	52.13±11.78	57.82±10.26
	S_{lms}+M	53.36±4.33	53.53±1.90	55.53±3.11	53.84±8.25	54.38±9.85	58.71±9.56
S1	F_0	32.02±10.16	49.11±7.04	55.47±7.60	52.85±5.05	54.07±10.64	52.33±13.70
S1+VT	F_0+S_{lms}+M	53.41±4.53	56.69±3.64	60.04±5.05	58.18±2.01	60.92±4.57	64.65±4.36
S2	S_{ilpr}	54.66±2.24	57.73±5.14	58.62±5.60	55.98±6.51	56.44±4.56	62.11±8.25
	S_{LP-gm}	53.40±4.36	53.54±5.22	55.40±6.61	53.91±8.56	50.37±6.71	49.91±6.51
	S_{ilpr}+S_{LP-gm}	58.18±3.30	59.40±5.01	59.62±6.24	58.36±6.95	57.53±5.54	62.43±7.74
S2+VT	S_{ilpr}+S_{lms}	56.60±1.76	58.57±3.25	59.46±3.97	58.78±7.32	58.33±4.11	65.04±5.54
	S_{LP-gm}+S_{lms}	54.37±4.48	54.86±5.02	56.69±6.07	54.98±8.17	53.36±9.97	58.82±8.44
	S_{ilpr}+S_{LP-gm}+S_{lms}	59.16±3.42	60.39±4.01	60.69±5.31	60.31±7.48	60.53±6.17	65.52±5.58
	S_{ilpr}+S_{LP-gm}+S_{lms}+M	59.33±2.44	60.77±1.38	61.46±4.22	60.52±7.16	61.05±5.64	66.07±5.95
S3	P1	41.27±9.66	45.98±7.98	52.20±8.14	52.38±10.36	55.42±6.48	53.51±9.46
	P2	52.44±9.50	55.89±10.75	63.77±5.39	57.27±8.41	60.75±9.40	59.26±12.04
	P1+P2	52.47±9.58	55.96±10.87	64.34±6.23	58.05±9.19	60.88±9.49	59.89±11.65
S3+VT	P1+P2++S_{lms}+M	58.34±4.40	60.56±5.47	65.57±5.06	57.84±6.87	59.52±9.42	62.92±7.07
	S1+S2+S3	60.49±3.64	63.06±5.27	67.91±7.70	60.38±9.65	62.46±9.81	66.07±6.98
	S1+S2+S3+S_{lms}+M	61.63±2.67	64.08±5.08	68.14±7.55	63.19±9.60	63.20±8.43	67.56±9.75

uses speech data from three folds for training and the remaining fold for testing. The training set is then divided in a 70 : 30 ratio to yield the training and validation sets. In order to see the effect of segment duration, the models are trained and tested on 1, 2, 3, 4, 5, and 6 second durations.

7.4 Results

To obtain the classification results, a four-fold speaker-independent cross-validation training strategy is employed for every combination. Table 7.1 shows the classification performance in average F1-score on 1, 2, 3, 4, 5, and 6 second durations. The table is

categorized into five broad parts: Vocal Tract (VT) features, tonal features as System 1 (S1), excitation source features as System 2 (S2), prosodic features as System 3 (S3), and finally the combined system (S1+S2+S3). The equations described in sub-section 5.7.4 and sub-section 6.6.4 are used to compute the combination of features from each individual system with the VT features at the score level. Furthermore, score level fusion is performed for the system combination (S1+S2+S3), as shown in equation 7.1.

$$S_{comb} = \alpha S_{S1} + \beta S_{S2} + (\gamma - \beta) S_{S3} \tag{7.1}$$

$$\gamma = 1 - \alpha \tag{7.2}$$

$$\alpha + \beta + \gamma = 1 \tag{7.3}$$

where S_{S1}, S_{S2}, and S_{S3} are the respective prediction scores from S1, S2, and S3. In equation 7.1, the α and β weight ranges from 0 to 1 and 0 to γ, respectively. The performance is calculated by varying these weight parameters in steps of 0.5. S_{comb} represents the combined score of the three systems. In addition, the combined systems are fused with the VT features to capture the complementary information. Equation 7.4 is used to fuse the VT features.

$$S_{VT} = \alpha_{VT} S_{S_{lms}} + (1 - \alpha_{VT}) S_{MFCC} \tag{7.4}$$

where, $S_{S_{lms}}$ and S_{MFCC} are the prediction scores generated by the S_{lms} and MFCC features, respectively. Finally, as shown in equation 7.5, the scores from the three systems (S_{comb}) are fused at score level with the scores from the VT features (S_{VT}) where, S_{all} is the combined score of the three systems with VT features (S1+S2+S3+ S_{lms}+MFCC).

$$S_{all} = \alpha_{all} S_{comb} + (1 - \alpha_{all}) S_{VT} \tag{7.5}$$

It is noticed from Table 7.1 that the highest performance in all categories is obtained when the segment duration is between 3 sec and 6 sec. The evaluated results show that the final system (S1+S2+S3+S_{lms}+MFCC) outperforms the individual systems (S1, S2, and S3) for all the segment duration. However, the final system achieves the best performance for the segment duration of 3 sec, with an average F1-score of 68.14%. An example of the predicted labels for the segment duration of 3 sec across the three dialects are shown in Figure 7-2 (a)-(c). The first row in Figure 7-2 (a)-(c) shows the detected speech regions marked in orange color. The second row in Figure 7-2 (a)-(c) shows the detected speech regions marked in orange color. The second row in Figure 7-2 (a)-(c) shows the predicted labels where, Changki dialect is labeled as 0, Mongsen as 1 and Chungli as 2.

Table 7.2: Detailed classification performance of Ao dialects at various segment duration. The results are reported in terms of mean (μ) and standard deviation (σ) for four-fold cross-validation. Vocal Tract, LMS, ILPR-LMS, LP-gammatonegram, Prosodic set 1, Prosodic set 2, System 1, System 2, System 3 are represented by VT, S_{lms}, S_{ilpr}, S_{LP-gm}, P1, P2, S1, S2 and S3, respectively.

	Features	Duration (in sec)	Accuracy ($\mu \pm \sigma$)	F1-score			
				Changki ($\mu \pm \sigma$)	Mongsen ($\mu \pm \sigma$)	Chungli ($\mu \pm \sigma$)	Average ($\mu \pm \sigma$)
Vocal Tract	MFCC (M)	1	49.44±5.78	38.20±20.54	38.27±11.59	62.04±6.42	46.17±5.24
		2	48.61±6.44	41.41±18.88	37.53±8.49	62.19±12.85	47.04±5.99
		3	51.89±4.48	38.62±22.84	43.06±12.37	65.76±9.93	49.14±4.69
		4	51.36±4.92	41.60±21.94	35.75±13.11	67.64±10.80	48.33±5.50
		5	51.51±4.15	39.81±22.61	42.90±12.27	67.80±13.93	50.17±4.25
		6	53.73±3.05	41.47±20.57	41.01±16.51	69.57±12.20	50.68±3.84
	S_{lms}	1	53.89±4.59	44.18±13.76	52.02±3.05	62.00±2.40	52.73±4.84
		2	53.96±1.88	42.78±9.20	52.47±7.25	63.88±5.92	53.04±1.75
		3	55.27±3.19	44.80±18.37	54.23±7.60	62.51±5.05	53.84±4.02
		4	55.29±7.46	41.80±10.59	54.00±17.14	62.77±13.33	52.86±7.99
		5	53.94±10.32	41.37±22.88	53.23±12.85	61.80±11.91	52.13±11.78
		6	60.61±8.59	52.27±21.19	56.61±16.59	64.59±14.31	57.82±10.26
	S_{lms}+M	1	54.84±4.25	44.47±14.10	51.58±3.70	64.02±2.35	53.36±4.33
		2	54.69±1.98	43.24±10.19	52.07±6.77	65.27±6.37	53.53±1.90
		3	57.10±2.63	46.66±17.04	54.24±7.17	65.71±4.33	55.53±3.11
		4	56.44±7.29	44.01±10.11	51.62±18.69	65.90±11.86	53.84±8.24
		5	55.85±8.59	44.84±18.50	52.82±12.17	65.49±11.75	54.38±9.85

		6	61.54±7.51	52.99±20.42	56.09±14.81	67.05±14.08	58.71±9.56
Tonal	F_0	1	38.36±6.23	21.77±17.84	25.00±14.77	49.30±5.40	32.02±10.16
		2	50.89±6.67	38.51±7.01	53.16±13.63	55.65±14.07	49.11±7.05
		3	55.87±7.23	44.28±13.07	62.32±3.52	59.81±12.80	55.47±7.61
		4	53.38±5.99	34.64±5.34	61.26±3.28	62.65±14.79	52.85±5.05
		5	54.33±10.77	42.77±11.23	62.04±6.53	57.39±19.38	54.07±10.64
		6	54.25±12.60	35.96±14.73	67.12±9.23	53.91±22.56	52.33±13.70
Tonal+VT	F_0+S_{lms}+M	1	54.98±4.38	44.26±14.36	51.64±3.86	64.35±1.95	53.41±4.53
		2	58.18±3.59	42.99±12.61	59.53±4.80	67.54±7.22	56.69±3.64
		3	61.16±5.08	47.20±15.79	61.55±2.78	71.36±7.99	60.04±5.05
		4	59.90±2.95	40.51±11.38	62.89±3.72	71.13±12.22	58.18±2.01
		5	61.79±4.90	47.46±12.36	63.85±5.20	71.45±11.26	60.92±4.57
		6	66.02±3.41	52.73±17.24	71.96±7.42	69.25±8.23	64.65±4.36
Excitation	S_{ilpr}	1	54.84±2.44	53.01±5.12	57.78±5.05	53.20±5.24	54.66±2.24
		2	58.51±4.67	55.37±8.09	62.13±8.15	55.69±15.90	57.73±5.14
		3	59.45±5.33	50.53±13.07	64.44±8.09	60.89±11.79	58.62±5.60
		4	57.32±4.48	55.87±6.57	58.42±13.19	53.64±13.26	55.98±6.51
		5	57.54±5.11	55.10±9.48	62.31±13.60	51.91±14.69	56.44±4.56
		6	64.55±7.27	62.42±12.06	68.40±8.21	55.51±24.27	62.11±8.25
	S_{LP-gm}	1	54.93±4.54	43.62±18.57	58.82±7.17	57.76±8.83	53.40±4.36
		2	55.01±5.59	41.73±18.51	60.59±9.63	58.29±11.31	53.54±5.22
		3	57.44±5.09	47.74±21.20	59.64±16.77	58.83±11.50	55.40±6.60
		4	57.08±7.49	46.43±23.90	55.56±26.09	59.73±12.80	53.91±8.55
		5	52.81±2.98	42.49±14.90	50.38±27.81	58.25±7.05	50.37±6.71
		6	53.03±3.68	40.90±19.52	51.06±23.71	57.77±12.91	49.91±6.51
	S_{ilpr}+S_{LP-gm}	1	58.84±3.40	53.15±11.69	62.14±6.90	59.27±8.73	58.18±3.30
		2	60.06±4.55	54.83±9.24	65.72±6.89	57.65±15.43	59.40±5.01
		3	60.66±5.60	51.14±15.94	65.59±9.65	62.13±12.50	59.62±6.24
		4	59.81±4.89	56.81±8.81	60.58±16.10	57.67±12.71	58.36±6.95
		5	58.43±5.62	54.95±10.29	62.16±16.47	55.48±12.16	57.53±5.53
		6	64.83±6.91	61.77±12.84	68.48±10.71	57.04±22.97	62.43±7.74
Excitation+VT	S_{LP-gm}+S_{lms}	1	55.90±4.73	43.95±18.66	59.38±6.16	59.77±7.99	54.37±4.48
		2	56.26±5.38	42.07±18.01	61.86±7.79	60.64±10.34	54.86±5.02
		3	58.45±4.98	47.40±20.67	60.09±14.46	62.57±11.01	56.69±6.07
		4	57.96±6.81	46.37±22.99	56.25±25.21	62.34±11.16	54.98±8.17
		5	55.25±9.10	43.33±21.20	54.26±15.01	62.50±10.06	53.36±9.97
		6	61.68±7.03	52.32±21.68	58.36±15.12	65.76±11.81	58.82±8.44
	S_{ilpr}+S_{lms}	1	56.67±1.79	53.48±6.01	58.75±4.06	57.56±4.39	56.60±1.76
		2	59.10±3.04	54.01±7.23	62.95±7.32	58.76±11.41	58.57±3.25
		3	60.46±3.68	50.79±13.35	64.40±7.54	63.18±11.21	59.46±3.97
		4	59.89±5.22	56.23±6.27	59.86±14.51	60.24±13.51	58.78±7.32

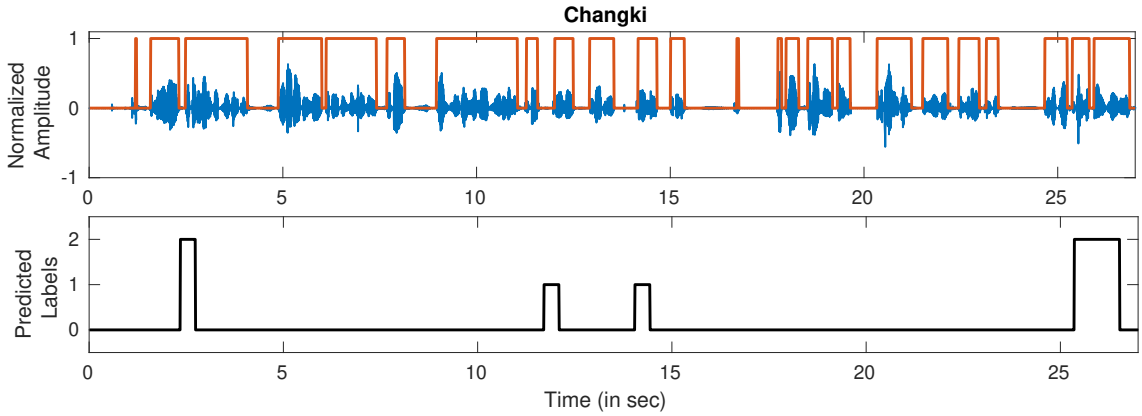
		5	59.20±4.84	54.79±11.07	63.53±12.77	56.66±10.43	58.33±4.11
		6	66.94±5.47	63.05±13.75	71.16±6.76	60.90±17.23	65.04±5.54
	S_{ilpr}+S_{LP-gm} +S_{lms}	1	59.86±3.68	52.31±13.11	62.18±4.12	62.98±6.67	59.16±3.42
		2	60.70±4.41	54.85±9.46	66.09±6.53	59.44±14.27	60.13±4.65
		3	61.87±4.69	51.85±16.27	66.01±9.03	64.20±12.38	60.69±5.31
		4	61.86±5.23	57.00±9.203	61.99±17.63	61.93±12.80	60.31±7.48
		5	61.22±5.66	54.73±12.32	63.23±14.72	63.63±8.79	60.53±6.17
		6	67.23±5.19	61.29±13.85	71.43±9.42	63.84±16.26	65.52±5.58
	S_{ilpr}+S_{LP-gm} +S_{lms}+M	1	60.27±3.11	51.85±12.59	60.31±4.19	65.83±5.77	59.33±2.44
		2	61.27±1.98	53.51±10.57	62.67±6.31	66.16±6.89	60.78±1.38
		3	62.67±3.94	52.33±13.41	64.52±9.65	67.54±10.67	61.46±4.22
		4	62.11±4.86	55.58±7.88	58.70±18.88	67.29±10.06	60.52±7.16
		5	61.80±4.69	54.79±10.67	62.22±14.12	66.13±9.14	61.05±5.64
		6	67.41±5.15	59.59±13.90	68.52±8.25	70.09±10.49	66.07±5.95
Prosodic	P1	1	43.12±7.98	37.14±8.02	47.38±16.76	39.28±20.99	41.27±9.66
		2	48.29±6.02	38.93±10.56	55.99±13.13	43.02±22.74	45.98±7.98
		3	53.39±6.48	44.84±12.69	60.60±14.25	51.18±15.96	52.20±8.14
		4	53.49±8.26	42.51±15.50	59.55±21.28	55.09±16.29	52.38±10.36
		5	55.81±5.20	46.72±11.21	65.27±12.74	54.27±18.29	55.42±6.48
		6	55.07±8.26	49.08±14.40	62.88±14.27	48.57±18.93	53.51±9.46
	P2	1	55.49±7.43	51.39±7.93	49.33±24.12	56.58±16.89	52.44±9.50
		2	58.57±8.22	55.33±4.14	55.02±26.63	57.34±17.54	55.89±10.75
		3	65.77±4.50	61.68±6.28	61.24±21.08	68.40±11.70	63.77±5.39
		4	60.81±6.27	52.33±8.14	58.76±29.06	60.71±21.21	57.27±8.41
		5	63.48±6.42	53.73±9.06	60.34±29.85	68.19±8.55	60.75±9.40
		6	62.27±9.02	55.83±5.66	63.63±31.18	58.33±19.47	59.26±12.04
	P1+P2	1	55.52±7.501	51.41±7.82	49.50±24.18	56.51±17.00	52.47±9.58
		2	58.62±8.33	55.10±4.11	55.45±26.54	57.33±17.80	55.96±10.87
		3	65.64±5.17	60.55±5.02	64.60±19.57	67.87±11.52	64.34±6.23
		4	61.37±6.93	52.94±7.44	59.49±29.07	61.71±21.13	58.05±9.19
		5	63.84±6.94	54.06±8.01	61.67±29.03	68.16±10.41	61.30±9.78
		6	62.19±9.65	55.02±10.58	66.13±25.51	58.52±20.18	59.89±11.65
Prosodic+VT	P1+P2+ S_{lms}+M	1	60.12±3.59	51.80±8.67	55.96±16.34	67.27±7.06	58.34±4.40
		2	61.37±5.52	52.85±8.77	59.95±13.34	68.89±5.15	60.56±5.47
		3	66.78±4.10	60.37±5.94	64.78±19.49	71.54±7.46	65.57±5.06
		4	61.07±4.90	48.45±10.92	57.22±28.43	67.86±11.00	57.84±6.87
		5	62.56±6.99	50.24±14.59	59.35±28.13	68.96±7.63	59.52±9.42
		6	64.97±5.69	56.58±17.95	67.92±13.75	64.27±9.28	62.92±7.07
S1+S2+S3	F_0+S_{ilpr}+ S_{LP-gm}+ P1+P2	1	61.44±3.48	54.80±11.19	64.25±9.65	62.43±10.58	60.49±3.64
		2	64.18±4.76	56.56±7.82	70.84±10.08	61.79±15.58	63.06±5.27
		3	69.01±6.59	62.01±7.88	70.81±18.05	70.91±12.08	67.91±7.70

		4	62.67±6.87	54.33±9.24	62.67±24.57	64.15±14.77	60.38±9.65
		5	65.41±6.46	54.67±7.76	63.36±30.22	69.34±11.03	62.46±9.81
		6	68.39±4.73	62.31±6.64	76.28±14.34	59.62±28.26	66.07±6.98
S1+S2+S3+VT	F_0+S_{ilpr}+ S_{LP-gm}+P1+ P2+S_{lms}+M	1	62.51±2.96	55.09±11.00	63.84±8.25	65.98±8.42	61.63±2.67
		2	64.79±5.18	56.22±9.73	69.80±8.25	66.21±12.65	64.08±5.08
		3	68.97±6.80	60.70±8.66	71.08±16.21	72.65±11.39	68.14±7.55
		4	65.25±7.10	55.43±7.73	62.20±24.32	71.94±10.75	63.19±9.60
		5	65.64±5.79	54.33±10.24	62.94±25.63	72.31±8.37	63.20±8.43
		6	69.27±8.33	63.00±10.24	72.61±20.22	67.06±13.40	67.56±9.75

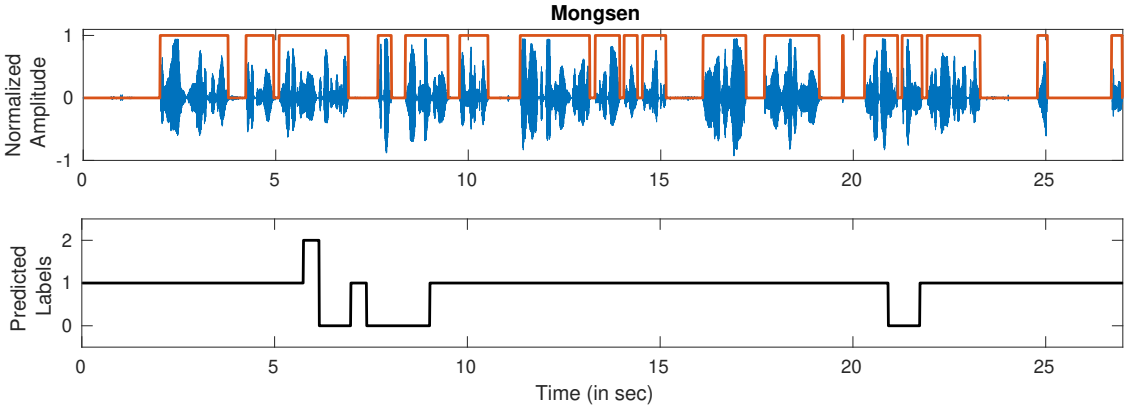
The detailed representation of the classification performance in the three dialects of Ao is reported in Table 7.2. It is seen that the proposed tonal, excitation and prosodic characteristics capture dialectal variations significantly, with only a marginal improvement observed when the proposed features are fused with the VT features. As a result, these findings support the effectiveness of the proposed features in classifying the three Ao dialects.

7.5 Application

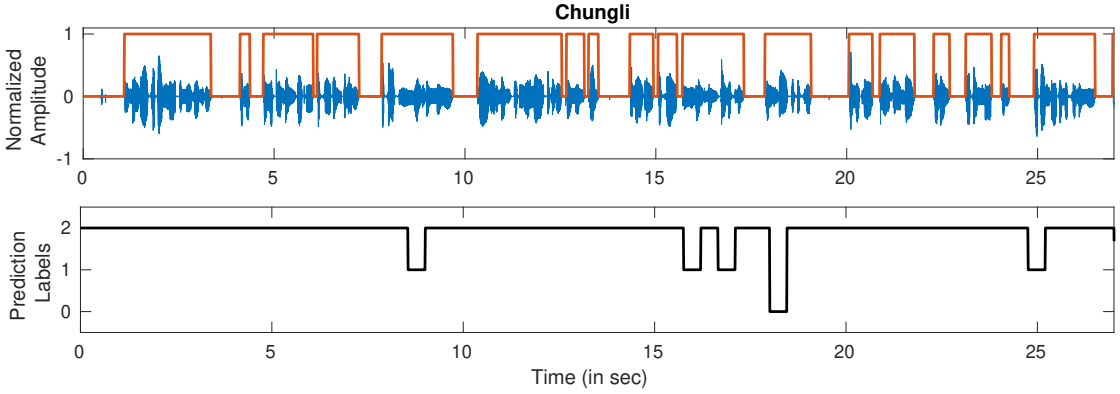
Automatic DID is important today due to its implication in ASR. The dialectal variations captured by automatic DID can be used in the pre-processing stage of an ASR system. In practise, if a speaker speaks for an automatic recognition of speech, the speaker will most likely utter a sentence or a small phrase (a group of few words) and then pause. The speaker, in particular, may use speech that is longer or shorter than 1 sec. A Breath Group (BrG) is composed of sentences that are separated by pauses. Therefore, as a specific use case, an experiment is carried out to evaluate the efficacy of approaches proposed for the Ao DID system at BrG level. To create the BrG corpus, the PasL corpus was annotated at BrG level with Praat 6.0.35 [68]. Hence, a total of 20466 tokens from the three Ao dialects was generated with the PasL corpus. Figure 7-3 shows the histogram plot for the 20466 tokens with respect



(a)



(b)



(c)

Figure 7-2: (a)-(c) Illustration of predicted labels for the best combination (S1+S2+S3+VT) in the three Ao dialects. First row: Speech signal with the detected speech regions marked in orange color for Changki, Mongsen, and Chungli dialects. Second row: Plot of predicted labels with Changki labeled as 0, Mongsen as 1 and Chungli as 2 for Changki, Mongsen and Chungli speech dialects

to time in seconds. It can be seen that maximum number of tokens is between $0.5-1$ sec segment duration. The BrG corpus was further augmented to include telephonic and two types of reverberated speech, as described in sub-section 5.6.3. In total, 81864 tokens were generated by combining the original and augmented speech data from the three dialects.

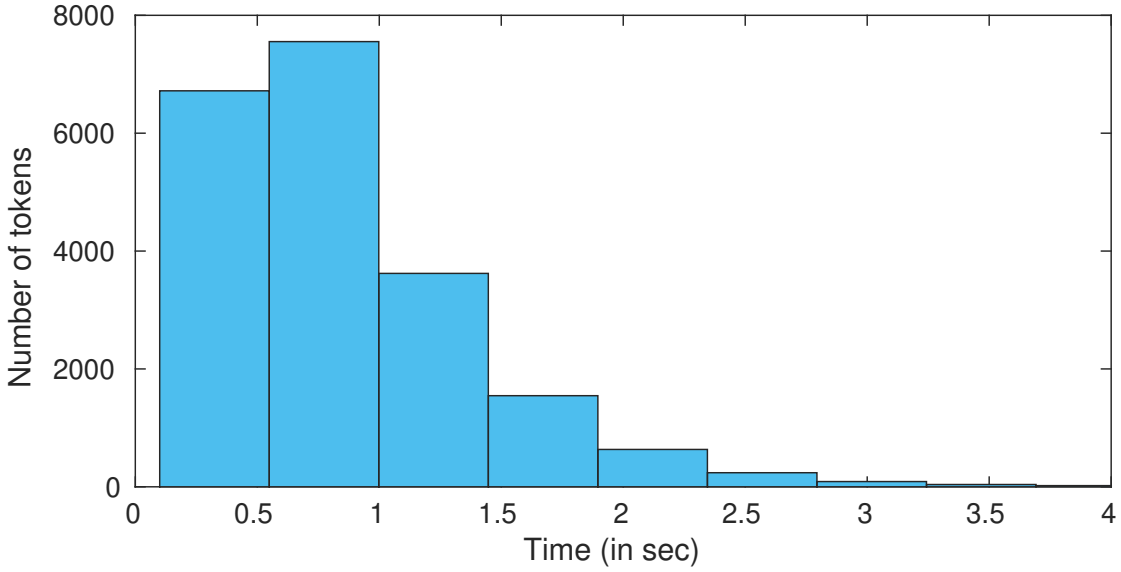


Figure 7-3: Histogram plot of BrG corpus for 20466 tokens

The experimental setup for evaluating the BrG corpus is the same as that shown in Figure 7-1 and discussed in section 7.3. A four-fold speaker-independent cross-validation training method is employed to obtain the classification results. By varying the frameshift, the frames for all of the features were kept constant at 300 for each token during feature extraction. For example, $S_{ilpr} \in \mathbb{R}^{40\times300}$ feature is fed into the classifier as an input, where 40 represents the feature dimension and 300 represents the time-axis. Table 7.3 shows the classification performance in the breath group database. The combination of features from each individual system with the VT features is computed at the score level using the equations presented in sub-section 5.7.4 and sub-section 6.6.4. Additionally, as indicated in equation 7.1, score level fusion

Table 7.3: Classification performance of Ao dialects in breath group. The results are reported in terms of mean (μ) and standard deviation (σ) for four-fold cross-validation. Vocal Tract, ILPR-LMS, LP-gammatonegram, LMS, Prosodic set 1, Prosodic set 2, System 1, System 2, System 3 are represented by VT, S_{ilpr}, S_{LP-gm}, S_{lms}, P1, P2, S1, S2 and S3, respectively. The best performance is highlighted and represented in bold.

			F1-score			
	Features	Accuracy	Changki	Mongsen	Chungli	Average
VT	MFCC (M)	46.93±5.07	36.44±9.33	44.21±13.18	56.75±11.77	45.80±4.55
	S_{lms}	54.86±2.96	42.81±11.78	56.71±5.48	60.45±9.08	53.32±3.59
	S_{lms}+M	55.21±2.84	42.92±11.84	56.84±6.38	61.02±9.76	53.59±3.38
S1	F_0	49.54±5.38	40.99±8.32	51.06±4.21	54.71±8.59	48.92±5.48
S1+VT	F_0+S_{lms}+M	58.22±1.56	45.26±9.82	60.30±6.22	64.49±9.98	56.68±1.17
S2	S_{ilpr}	49.26±4.92	43.59±6.33	50.86±7.78	50.37±11.41	48.27±5.21
	S_{LP-gm}	50.60±3.52	46.25±9.62	53.46±11.19	48.20±6.70	49.30±4.17
	S_{ilpr}+S_{LP-gm}	52.89±4.31	47.18±8.45	55.87±10.33	51.37±10.34	51.48±5.14
S2+VT	S_{ilpr}+S_{lms}	56.35±2.79	45.47±11.43	58.78±5.11	60.05±9.26	54.77±3.33
	S_{LP-gm}+S_{lms}	56.41±2.66	47.60±10.02	59.02±7.86	58.13±9.28	54.92±3.33
	S_{ilpr}+S_{LP-gm}+S_{lms}	57.07±2.76	46.07±11.67	59.40±5.66	60.87±9.52	55.45±3.38
	S_{ilpr}+S_{LP-gm}+S_{lms}+M	57.09±2.69	45.90±11.58	59.37±5.91	61.08±9.56	55.45±3.30
S3	P1	43.96±3.50	37.22±10.41	49.18±12.41	40.12±15.73	42.18±4.10
	P2	50.50±6.93	44.52±7.89	49.52±19.29	53.83±10.24	49.29±7.62
	P1+P2	51.14±6.66	44.97±8.17	50.81±19.38	53.72±11.68	49.84±7.46
S3+VT	P1+P2+S_{lms}+M	57.55±4.63	46.04±11.33	58.62±12.95	62.68±11.07	55.78±5.31
	S1+S2+S3	57.12±5.77	49.29±8.41	61.58±11.87	56.86±13.46	55.91±6.28
	S1+S2+S3+S_{lms}+M	60.27±3.98	50.08±10.26	63.07±8.85	63.06±11.83	58.74±4.61

is carried out for the system combination (S1+S2+S3). Finally, equation 7.4, and equation 7.5 are used to compute the final system score (S1+S2+S3+S_{lms}+MFCC). It is noticed from the table that a decent performance is achieved by acquiring an average F1-score of 58.74% for the final system, which is slightly comparable to the 1 sec performance in Table 7.2.

7.6 Summary

A combined Ao DID system with tonal, excitation, and prosodic features are proposed, consisting of three DID modules explored in previous chapters (chapter 4,

5, and 6). The attention-based Bi-GRU classifier is used to model F_0 and prosodic features. While the attention-based CNN-BiGRU classifier is used to model the excitation and baseline features. The experiment is conducted at various segment durations of 1, 2, 3, 4, 5, and 6 secs. The score level fusion is used to combine the three modules. When the proposed features are combined with the VT features, the best performance is obtained. Also, the 3 sec segment duration provides the best overall performance for the system. When it comes to practical application, however, a speaker may speak for less than 1 sec. As a result, in order to evaluate such a scenario, an experiment is carried out in which the PasL corpus is annotated at Breath Group (BrG). It can be seen that the majority of the tokens in the BrG corpus fall between $0.5 - 1$ sec. The BrG corpus is evaluated using the same experimental setup as the combined Ao DID system. The result with the BrG corpus showed comparable performances to the final Ao DID system with 1 sec segment duration.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 8

Conclusions

Overview

This chapter provides the summary of the work presented in this dissertation for automatic DID in Ao. Based on the contributions of the thesis, some possible future directions are also discussed.

8.1 Summary

This dissertation aims to identify the three dialects of Ao, a Tibeto-Burman language spoken in the North Eastern part of India, Nagaland. The Ao community is one of the major tribes and has the maximum population among the various Naga tribes. The natives are primarily settled in the Mokokchung district. The language is divided into three distinct dialects, namely, Chungli, Mongsen and Changki. Ao is a tonal language that has three lexical tones, High (H), Mid (M) and Low (L). The three dialects differ in terms of tones for the same word. As Ao is an under-resourced language, only a handful of work in the field of linguistics is available, and no work in automatic method is known to us. Several features are proposed in this thesis based on the differences in tone across the three dialects. The contributions incorporated for this dissertation are summarized below.

(i) **Changki Mongsen Chungli-Ao (CMC-Ao) Corpora:** As Ao is a language with inadequate resources, there are no publicly available speech resources for Ao. As such, the first contribution of this thesis is in the collection and annotation of speech data with a total of 71 unique native speakers from the three Ao dialects. For this research work, Changki, Khensa and Mopungchuket villages were chosen to represent Changki, Mongsen and Chungli dialects respectively. The CMC-Ao corpora consisted of three types of speech corpus: disyllabic words (DiW), trisyllabic words (TriW), and passage level (PasL). The DiW corpus was recorded from 30 native speakers in total, comprising target words with six disyl-

labic minimal sets. The TriW corpus consisted of 40 trisyllabic target words recorded from 36 native speakers in total. The PasL corpus was recorded from 24 native speakers for the three dialects. The Bible passage, "The parable of the prodigal son" was read by each speaker in four sessions. The total duration of the PasL corpus was approximately six hours, consisting of 96 passages in total. The speech data for DiW and TriW corpora were annotated and tone boundaries were marked manually using Praat 6.0.35.

(ii) **Tonal Feature Based Dialect Identification:** This thesis investigates the acoustic analysis of tones in the three Ao dialects using DiW corpus. The acoustic properties of the tones in the Chungli and Mongsen dialects are noticeably different from those in the Changki dialect. As a result, automatic DID in the three dialects of Ao is investigated using F_0 features along with its Δ and $\Delta\Delta$ derivatives in TriW corpus. The F_0 features are extracted using two methods: ZFF and Praat method based autocorrelation. The results showed that ZFF approach performed better in the Ao DID system. The ZFF method can reliably and effectively capture the micro-prosodic information of F_0 contour as it is extracted from the GCI location. Statistical analysis also shows that Ao dialects can be predicted more accurately using ZFF-derived F_0 features. Hence, an automatic Ao DID is built using MFCC and SDC features with the inclusion of F_0 features, confirming the significance of dialect-specific information.

(iii) **Excitation Source Feature Based Dialect Identification:** Following that, excitation source features are explored as there are no source features used in DID tasks to the best of our knowledge. Initially, the TriW corpus is used to design an Ao DID system with the RMFCC feature, yielding a fair result. The RMFCC features are then combined with the baseline MFCC and SDC features, giving improved performance. Further, an approximate representa-

tion of source features is proposed using ILPR-LMS and LP-gammatonegram. Using the PasL corpus, an attention-based CNN-BiGRU model is proposed to evaluate the performance of the Ao DID system within a broader framework. The attention-based CNN-BiGRU classifier learns the features automatically from the time-frequency representation of excitation source features. MFCC and LMS are used as baseline features to capture VT information. The experiments are divided into several categories. First, the original PasL corpus is used to identify the three Ao dialects. The result showed decent performance when the source features are fused with the VT features. Secondly, due to the low performance, data augmentation is carried out. Data augmentation increased the combined source and VT feature performance by 14%. Third, the proposed architecture is subjected to hyper-parameter tuning, which improves the average F1-score by 3%. Finally, at various segment durations, the effect of segment duration is investigated using a human perception test and an automatic DID method. For the combination of source and VT features, the final system improves performance by 61.46%.

(iv) **Prosodic Feature Based Dialect Identification:** Prosodic features for the Ao DID tasks are explored based on the type and assignment of tones in the three dialects. The statistical prosodic features are initially applied to the automatic Ao DID task with the TriW corpus. The prosodic features are classified into three categories: F_0ST, loudness, and VQT. As the baseline feature, statistical MFCC features are used. The best results are obtained when the prosodic features are combined with MFCC. The results also show that VQT (voice quality and temporal) plays an important role in distinguishing the three Ao dialects. As a result, attention-based Bi-GRU is proposed to capture the temporal information of the three dialects. Using the PasL corpus, this architecture is proposed to design a generalized Ao DID system. For classification, the LLD

prosodic features P1 and P2 are used, with MFCC-LLD as the baseline feature. When combined with MFCC-LLD, the prosodic features achieve the highest performance of 64.46% average F1-score.

(v) **Combined Framework for Dialect Identification in Ao:** After exploring the significance of tonal, excitation and prosodic features in various Ao DID modules, a combined Ao DID system with the three modules is proposed. TriW corpus was used in the previous experiment with F_0 features. The PasL corpus, on the other hand, is used for all three modules in the combined system. The attention-based Bi-GRU classifier is used to model F_0 and prosodic features. While, attention-based CNN-BiGRU is used to model the excitation source features. MFCC and LMS are used as the baseline features modeled with attention-based CNN-BiGRU. To combine the three systems, score fusion is computed. The experiment is carried out at various segment durations. For all durations, the best performance is obtained when the proposed features are combined with the VT features. Furthermore, the PasL corpus is annotated at Breath Group (BrG) to evaluate the effectiveness of the proposed features from the point of application. The final system performed decently in identifying the three Ao dialects using the BrG corpus.

8.2 Contributions of the thesis

The major contributions of the work done in this dissertation are as follows.

(i) The Changki Mongsen Chungli-Ao (CMC-Ao) Corpora was specifically developed for this dissertation to aid automatic DID in Ao. The DiW corpus and TriW corpus in the CMC-Ao corpora are manually annotated for target words with tone boundaries.

(ii) Acoustic study is conducted for the three dialects of Ao. Tone dynamics in the three dialects are also investigated.

(iii) F_0 features along with its Δ and $\Delta\Delta$ derivatives are proposed for automatic Ao DID task.

(iv) Excitation source feature such as RMFCC is proposed to build an Ao DID system. In addition, an approximate representation of source features such as ILPR-LMS and LP-gammatonegram are proposed.

(v) An attention-based CNN-BiGRU is proposed for automatic feature learning of the time-frequency representation of source features.

(vi) Prosodic features extracted from openSMILE toolkit are investigated for identifying the three Ao dialects.

(vii) To learn the temporal characteristics of the prosodic features, an attention-based Bi-GRU is proposed.

(viii) A combined Ao DID system based on the tonal, excitation, and prosodic feature modules is proposed.

(ix) BrG corpus is developed to assess the efficacy of the combined Ao DID system as a specific use case.

8.3 Direction for future work

This dissertation investigates various aspects of an automatic Ao DID system. Some extensions to this work are proposed based on the findings of these investigations. The following are some potential future directions.

- **Expansion of the CMC-Ao corpora:** Chungli is known as the standard dialect of the language. Therefore, the written form is only available in Chungli.

Furthermore, the collected speech data was in read speech. As a result, the proficiency of the speakers was not taken into account when collecting speech data for this thesis. Hence, CMC-Ao corpora can be expanded to include spontaneous and conversational speech data in the future by considering the proficiency in the language.

- **Speaker-wise experiment:** This work can be expanded by evaluating the speakers individually to see how they performed in the DID task. There may be instances where the overall performance decreases as a result of one or more speakers. Hence, each speaker can be subjected to a thorough analysis in the future.

- **Effect of gender:** In male and female speech, the fundamental frequency (F_0) and its corresponding harmonics have quite different ranges. These variations in characteristics could have an impact on the classification task. In addition, Ao being a tonal language, the tone information of speech signal may change according to different gender. Hence, effect of gender for DID task is worth exploring in the future.

- **Develop DID system in other languages:** To the best of our knowledge, this is the first thesis to devise an automatic DID system among the various languages of Nagaland. Other languages, such as Angami, Sumi and Lotha are also majorly spoken in other parts of Nagaland. Therefore, dedicated DID systems for these languages need to be developed in future.

- **Incorporation of DID system for ASR:** In the future, the Ao DID system could serve as a pre-processing step for a higher-level application such as Automatic Speech Recognition (ASR). Hence, the proposed DID system can be integrated into an ASR system to improve speech recognition.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Appendix A

Materials

A.1 Disyllabic words (DiW)

Table A.1 shows the six minimal sets with only two sets (/metsü/ and /temang/) common for the three dialects of Ao. However, it is to be noted that the common sets have varying number of examples for each dialect.

Table A.1: Disyllabic word list.

	Gloss		
Disyllabic words	Changki	Mongsen	Chungli
/mesep/	-	-	'got drunk'
	-	-	'kiss'
	-	-	'stinging feeling'
	-	-	'suckle'
/metsü/	'deer'	'deer'	-
	'kick'	'kick'	'kick'
	-	-	'rearing'
	'salt'	'salt'	'salt'
	'saliva'	'saliva'	'saliva'

	-	-	‘seeds’
/mokba/	‘grinding rice’	-	-
	‘hatching eggs’	-	-
	‘wrap a shawl’	-	-
/phuba/	‘animal giving birth’	-	-
	‘carrying load’	-	-
	‘blowing’	-	-
/senba/	‘cooking in bamboo’	-	-
	‘leaking’	-	-
	‘procession’	-	-
	‘speak ill of somebody’	-	-
/shishi/	-	-	‘to buy’
	-	-	‘get up’
	-	-	‘to tell’
/tanük/	-	-	‘nephew’
	-	-	‘small bits’
	-	-	‘smaller denomination’
/tasa/	-	-	‘acrid odour’
	-	-	‘aging’
	-	-	‘high tune’
	-	-	‘plan’
/techa/	-	‘don’t do it’	-
	-	‘hot’	-
	-	‘ribs’	-
	-	‘wings’	-
/tekong/	-	‘dry’	-
	-	‘neck’	-

	-	‘valley’	-
/telang/	-	‘bottom’	-
	-	‘buttock’	-
	-	‘far’	-
/teli/	-	‘don’t buy’	-
	‘move away’	‘move away’	-
	‘veins’	‘veins’	-
	‘vine’	-	-
/temang/	-	‘all’	-
	‘belief;believe’	‘belief;believe’	-
	‘body’	‘body’	‘body’
		-‘dark’	-

A.2 Trisyllabic words (TriW)

The TriW corpus consists of 40 trisyllabic utterances with the same meaning for the three Ao dialects listed in Table A.2. As shown in the table, the 40 words are common for the Changki and Mongsen dialects. However, there are lexical differences for some words in the Chungli dialect (highlighted in blue color).

Table A.2: Trisyllabic word list.

Changki | Mongsen | Chungli	Gloss
baksaba | baksaba | bakshiba	‘to scatter’
baktila | baktila | bakzüba	‘group of people moving abreast’
benjaba | benjaba | benjaba	‘to disturb; to create confusion’
bentangba | bentangba | bentangba	‘a peacemaker; to stop a quarrel/fight’
bowaba | bowaba | bowaba	‘to charm; to attract’

changaba | changaba | yintetba	‘growing up of a boy’
changati | changati | asangur	‘young unmarried men; teenage boys’
changremri | changremri | sangremer	‘windower’
chetchaba | chetchaba | azhitep	‘a slight touch’
endangba | endangba | nukshiba	‘to advance one’s attention towards someone’
ingsaba | ingsaba | epshiba	‘to scatter dry leaves’
itepba | itepba | azuitba	‘embrace’
jangtsülong | jangtsülong | tsungchilong	‘a stepping stone’
jemrepba | jemrepba | ajemrep	‘many people crying together’
jenchetba | jenchetba | jentetba	‘to crush by stepping o’
jeptepba | jeptepba | ajeba	‘able to flee’
lazadi | lazadi | ayirla	‘young marriageable girl’
molutzü | molutzü | molutzü	‘ocean/sea’
mongliba | mongliba | mongshiba	‘to mix clothes or any leafy things in a container’
nemetba | nemetba | nemetba	‘laying of hands on a patients head’
oleplang | oleplang | olepdang	‘response’
ongmetang | ongmetang | aongmetang	‘echo’
phiphike | phiphike | züzüa	‘feeling of numbness’
pilaba | pilaba | pilaba	‘to separate; to part’
tedinu | tedinu | kudinu	‘her cousin sister’
tekaksa | tekaksa | tekaksa	‘broken pieces’
tekülem | tekülem | tekülem	‘worship’
tekolok | tekolok | tekolok	‘brain’
teluba | teluba | tanüngba	‘remainder; leftover’
telumi | telumi | telumi	‘starvation’
temeka | temeka | takumret	‘jaw’
temelong | temelong | temelong	‘heart’

temesen | temesen | temesen	'liver'
tephela | tephela | tephela	'navel'
teyimla | teyimla | teyimla	'hope'
tsüksaba | tsüksaba | tsüksaba	'to crush into pieces'
tsümiba | tsümiba | penbuba	'wanting to have'
tsüpaba | tsüpaba | tsübuba	'fear; afraid'
wamaba | wamaba | wamaba	'to slice into pieces'
watangba | watangba | watangba	'to saw/cut into two pieces'

A.3 Passage level (PasL)

The written form of the Bible passage, "The parable of the prodigal son" was available only in the Chungli dialect. As a result, the Changki and Mongsen dialects were translated. The Chungli text, translations to Changki and Mongsen dialects, and the English meaning are provided below:

(i) **Chungli:**

Nisung ka jabaso ana liasü. Tanubusang jagi tebu dang ashi, "Oba, ne rongsen nungi kü shilem kechi lir aji kü nem kua." Anungji tebui pei rongsen parnok nem lemsa agütsü. Anogo ishika lir külen jabaso tanubusangi pei oset ajak bener alima talang kati aeni ao. Idakji pai pei rongsen lalushibonga temenen benshia ali. Aser pai pei rongsen ajak endokmar külen iba lima nungji aya-wara kanga tulu ka adok aser pa kanga sensak aten. Idangji iba lima alir nübur rongnung ka den pai semloka ali, aser kibur jagi moapu nung ak anüktsü yok. Aser ak-i achiba matsüklashi tekep achitsü pai yongya saka shingaia pa nem kecha magütsü. Saka kodang pa temulong tanga aru pai ashi, Kübu den ayangertem kwika chiyungtsü peria aser ano talia angur, saka nibo idaki taya agi tasür. Ni apusoa kübu dangi odi, aser ni pa dang ashitsü, "Oba, kotak anema aser ne madang

ni temenen menaogo; ni joko ne jabaso ta ajatsü metemsü; ne ayangertem rongnung ka ama ni amshiang."

Aser pai apusoa tebu dangi aru. Saka pa ano apiga nung alidang, tebui pa ngua aochia bilema asema ao aser tekong nung azuit aser pa tebang mesep. Idangji techiri tebu dang ashi, "Oba, kotak anema aser ne madang ni tai menaogo; ni joko ner jabaso ta ajatsü metemsü." Saka tebui kilirtem dang ashi, "Sobutsü tajungtiba yakta bena arung aser aji pa nem sobujang, pa temeyong nung küri aser tetsüng nung tsüngsem semokjang. Aser nashi chanu teyirabaji anir arung aser tepsetang aser asenoki chiyunga pelatepdi. Kechiaser kü jabaso ibai tasü liasü, aser tanaben taküm lir; pa sama-a liasü saka tang meyipa nguogo." Aser parnoki pelateptsü tensük.

Idangji par jabaso tambusangji alu nung liasü; aser pai adoka ki anasa arudang kentenba aser tsüngsangba ola pai angashi. Aser pai kilirtem rongnungi ka aja aser ya kechi tetezü ta asüngdang. Aser kilir jagi pa dang langzü, "Nenu shilanga lir, aser nebui nashi chanu teyirabato tepset kechiaser pai nenu junga aser anema lia meyipa angu." Saka pa tediji ain adok aser kidangi metuli ta ashi. Anungji tebui kimai adoka pa dang mepishi. Saka pai tebu dang langzü, "Ajiangjo, küm paikati ni ne dang tenzüka aru, aser nai ashiba o kodanga mangai mali; saka kü tembartem den pelateptsü asoshi kü nem nai kodanga nabong chanu ka danga magütsü. Saka kodang iba ner jabaso shibai nüktapangtatsürtem den ne sen chima aser kodang pa meyipa aru, pa asoshi nai nashi chanu teyirabato tepset." Idangji tebui pa dang langzü, "Kü jabaso, na teti kü den lir aser kü meyong ajakji ne meyong lir. Saka asenoki temulong chia pelateptsü tim dang lir, kechiaser nenu tasü liasü, aser tang pa tanaben taküm lir, pa madoka liasü saka tang pa nguogo."

(ii) **Changki translation:**

Ami a jabaso anet lichakü. Nozaba tsünü teba dangko sakü, "Abao, nang rongchen mino ni salem koba kalü ibatsü kü li koti." Tetsü tebanü papa rongchen tongi li lemsa kikü. Cheniko ichalaka liri senko jabaso nozabachangnü pa oset tamang eniri alima telanganü aing wakü. Ibatsüko panü pa rongchen lalüshibongkü temenen enlikü likü. Tsütechairi panü rongchen tamang jokchemiri senko iba lima tsüko ayim-wara samdang tepeti a tsüka tsütechairi pa samdang süngja kümkü. Tsütechatemko iba lima ko liri mijang rongko a no pa chenokü li, tsütechairi kibuba tsünü moapü ko aok nakiba zükü. Tsütechairi aok nü chaba napakosü tekep chaiba pa nü zongza, tako sünükü pa li takü mekilü. Tako kotem pa temolong zangkü rakü pa nü sakü, Aba no yayangrilong koza ching-i perikü tsütechairi ano talila ongri, tako nilü ibiko changti nü tesürio. Ni kepkü aba mang nü waro, tsütechairi ni pa dang ko sai, "Abao, kotak anemkü tsütechairi nang madangko ni temenen menaogo; ni joko nang jabaso ta chaiba metemsüa; nang yayangri rongko a asa ni enliong."

Tsütechairi pa nü kepkü teba mangnü rakü. Tako pa apikü litem ko, teba nü ongkü aochikü bilemkü semkü wa tsütechairi tekong ko yitkü tsütechairi pa tebang ko mechepkü. Tsütechatemko techarinü teba dangko sakü, "Abao, kotak anemkü tsütechairi nang madangko ni temenen menaogo; ni joko ning jabaso ta chaiba metemsüa." Tako tebanü kiliri dangko sakü, "Sübui tepongba zakte eni rong tsütechairi ibatsü pa li süboiyang, pa temeyong ko kori tsütechairi techang ko changjem chemokiang. Tsütechairi masü teza techaningbatsü noiri rong tsütechairi kosetang tsütechairi esünglanü chingkü pelatepro. Talitalü ni jabaso ibüi tesü lichakü, tsütechairi nebenroba teküm liano, pa chamakü lichakü, tako toko meyipkü ongoko." Tsütechairi tongla pelatepiba tenlakü.

Tsütechatemtsüko pari jabaso tezenbachangtsü alu ko lichakü, tsütechairi pa tsükakü aki anacha ratemko aken tenba tsütechairi tsüngsangba ola panü zakü.

Tsütechairi panü kiliri rongko a jatakü tsütechairi iba ta tetezü te süngchakü. Tsütechairi kiliri tsünü pa dangko langlikü, "Nenu sülangkü liano, tsütechairi nebanü masü teza techaningbatsü kosetano talitalü panü nenu pongkü techairi anemkü likü meipkü ongkü." Tako pa tengatsü azen tsüka tsütechairi akinü meküra te sakü. Tsütechabanü tebanü kimanü tsükakü pa dangko mepisükü. Tako panü teba dangko langlikü, "Etsütsüang, aküm pazayai ni nang dang tenlakü rano, tsütechairi nangnü saba ayo kotemkü metsülü melia; tako ni atsüri no pelatepiba atemkü kü li nang nü kotemkü nabong teza a dangkü mekitsüa. Tako kotem iba ning jabaso sünü niktapangta anoti no nang achen alangkü tsütechairi kodem pa meipkü ralü, pa atemkü nang nü masü teza techaningbatsü kosetkü." Tsütechatemko tebanü pa dangko langlikü, "Kü chari, nang teti ni no likü tsütechairi kengdang melong tamangtsü nang endang melungo. Tako esüngla temelong chakü pelatepibatsü tim dango, talitalü nenu tesü lichano, tsütechairi pa nebenroba teküm liano, pa mairi lichano tako toko pa ongoko."

(iii) **Mongsen translation:**

Ami a anu anet licha. Tonget jenko nuzaba chang nü tepa dangko sa, "Apa, nü chenmang bhinü kü salem khiang." Taiko tepa sünü ano anet süli pa chenmang lemsakü khi. Tsüngi ko echa lir nuzaba chang sünü pa oset temang hernemer alima telang anü ahing wa. Iba lima süko pa khaba temang mashi hemshi ka temenen mapa chining. Iba mapang tsüko aya-wara tepetia tsüka techar pa bendakba tia nü za. Jokola pa sü ami ar ki ko kili ka li. kifur sünü aok nakro moapu nü zükya. Pa nü aok li khiba napakbi tekep tsamika atsü. Tepakoka sü-nü pa-li tsongvi mekhila. Pa joko tengachetba-a nü ra, tebar ki ko kilir long tsongbiteperi hongaka liba bilemchet. Pa joko yamüsoko tepa dangnü war, "Apa, ni nü kotak kha nang net anemaka tai menaogo; ni nü tsar chaviba malanglao; taiko neng ki ko kilir a dangka chakü limiro" te

saviba bilemcheter yamüso.

Aki tongviba telanga liga, tepa ne pa honger semaka wa, techar tekong ko eiter tebangko mejeb. Iba mapang ko tetsa nü sa, "Apa, ni nü tsar chaviba melanglao." Tepakoka tepa nü kilirlong dangko sa, "Sobovi tarutiba zakte bener rar kü tsa sobubiang, pa temeyong ko küri, tejang ko jangchem jemokbiang, masü teza tametba henseter tsongtepaka pelatepro" te ayonglak. Ano tepa sünü sa, "Kü tsar-i süjokoko te lija, takola teküm kümogo, mhar lija tako meyipaka hongoko". Techar tongla pelatepaka benjongtep.

Iba mapang tsüko par tsar tezümba chang alu bhinü tetsükarba par ki ko achu tentepbaka yartepba ya. Pa nü kilir a tsar, "Ibi jeba char" te semtzü. Kilir sünü langli, "Nenu meyipaka rar lio, nüba nü masü teza tametba sü nenu arukadang aki nü mangepbaka raba tening ko hensetogo." Patsü yar teti sü racha ka aki nü mekoabio tesa. Tepa kima nü tzükar pa dangko mepishiküsa. Tepakoka pa nü langli, "Atsüangma, aküm payai ni nü nang tenlakaka ra, nang nü saba yu koyim ka münghala melika; tebakoka kü metemerlong den pelatepbiba atemaka küli napong teza a dangka mekhitsüla. Tani nü tsar ibai sübanü niktapangtaka nü rongchen tsamar meyipaka ra, pa atemaka masü teza tametba henset." Iba mapang ko tepa nü langli, "Kü tsar, nang teti kü den lir, kü khet ko liba temang nü indang, esa tani pelatepbiba mapang shitak-o, jepacha tepala nenu süyokogo te lija, tepakoka mesüla meyipbaka rar li, pa mhar lija, tako hongogo."

(iv) **English:**

There was a man who had two sons. The younger one said to his father, "Father, give me my share of the estate." So he divided his property between them. Not long after that, the younger son got together all he had, set off for a distant country and there squandered his wealth in wild living. After he had spent everything, there was a severe famine in that whole country, and he began to

be in need. So he went and hired himself out to a citizen of that country, who sent him to his fields to feed pigs. He longed to fill his stomach with the pods that the pigs were eating, but no one gave him anything. When he came to his senses, he said, "How many of my father's hired servants have food to spare, and here I am starving to death. I will set out and go back to my father and say to him, Father, I have sinned against heaven and against you. I am no longer worthy to be called your son; make me like one of your hired servants."

So he got up and went to his father. But while he was still a long way off, his father saw him and was filled with compassion for him; he ran to his son, threw his arms around him and kissed him. The son said to him, "Father, I have sinned against heaven and against you. I am no longer worthy to be called your son." But the father said to his servants, "Quick! Bring the best robe and put it on him. Put a ring on his finger and sandals on his feet. Bring the fattened calf and kill it. Let's have a feast and celebrate. For this son of mine was dead and is alive again; he was lost and is found." So they began to celebrate.

Meanwhile, the older son was in the field. When he came near the house, he heard music and dancing. So he called one of the servants and asked him what was going on. "Your brother has come," he replied, "and your father has killed the fattened calf because he has him back safe and sound." The older brother became angry and refused to go in. So his father went out and pleaded with him. But he answered his father, "Look! All these years I've been slaving for you and never disobeyed your orders. Yet you never gave me even a young goat so I could celebrate with my friends. But when this son of yours who has squandered your property with prostitutes comes home, you kill the fattened calf for him!" "My son," the father said, "you are always with me, and everything I have is yours. But we had to celebrate and be glad, because this brother of yours was dead and is alive again; he was lost and is found."

Appendix B

Background information of participants

Table B.1: Information of participants for disyllabic words (DiW)

Name	Age	Sex	Education	Language known	Mother tongue
WP	27	F	Post Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
ST	45	F	HSSLC	Changki, Chungli, Mongsen, English, Nagamese	Changki
LN	25	F	Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
AT	58	F	HSLC	Changki, Chungli, Mongsen, English, Nagamese	Changki
MM	56	F	HSLC	Changki, Chungli, Mongsen, English, Nagamese	Changki
AR	42	F	HSSLC	Changki, Chungli, Mongsen, English, Nagamese	Changki
AM	24	F	Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
TS	38	F	Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
ST	33	M	Post Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AT	35	M	PhD	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
NR	31	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
TI	30	F	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AS	30	F	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AM	32	F	PhD	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
KT	33	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen

SL	33	M	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AN	33	M	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
MZ	37	M	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
MT	40	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
KC	31	M	Post Grad.	Mongsen, English, Hindi, Nagamese	Mongsen
AJ	35	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AC	47	F	HSLC	Chungli, Nagamese	Chungli
AK	25	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AS	31	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AL	50	F	HSLC	Chungli, Nagamese	Chungli
TR	32	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
KR	57	M	HSSLC	Chungli, Nagamese	Chungli
TS	45	M	Grad.	Chungli, English, Hindi, Nagamese	Chungli
LM	34	M	Grad.	Chungli, English, Hindi, Nagamese	Chungli
CS	58	M	Grad.	Chungli, English, Nagamese	Chungli

Table B.2: Information of participants for trisyllabic words (TriW)

Name	Age	Sex	Education	Language known	Mother tongue
TS	39	F	Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
CB	44	F	Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
WT	43	F	Post Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
HL	38	F	Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AR	26	F	Post Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
AT	36	F	Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
ST	34	M	Post Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AT	36	M	PhD	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AM	45	M	HSSLC	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
TL	41	M	HSSLC	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AM	37	M	Post Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AR	27	M	Post Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
AM	33	F	PhD	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen

NR	32	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AN	26	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
ST	32	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AJ	31	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
ST	31	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AK	31	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
IM	24	M	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AL	27	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
IM	33	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
IT	34	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
IL	39	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AR	26	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
IM	31	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AK	26	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
IM	57	F	HSLC	Chungli, English, Nagamese	Chungli
IS	30	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AR	40	F	Grad.	Chungli, English, Hindi, Nagamese	Chungli
AH	45	M	Grad.	Chungli, English, Nagamese	Chungli
IB	35	M	Grad.	Chungli, English, Hindi, Nagamese	Chungli
IN	25	M	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AB	25	M	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
ZL	28	M	Grad.	Chungli, English, Hindi, Nagamese	Chungli
AC	30	M	Grad.	Chungli, English, Hindi, Nagamese	Chungli

Table B.3: Information of participants for passage level (PasL)

Name	Age	Sex	Education	Language known	Mother tongue
TS	40	F	Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AR	27	F	Post Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
AT	60	F	HSLC	Changki, Chungli, Mongsen, English, Nagamese	Changki
MD	33	F	Grad.	Changki, Chungli, Mongsen, English, Nagamese	Changki
ST	35	M	Post Grad.	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki

AT	37	M	PhD	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
NS	43	M	HSSLC	Changki, Chungli, Mongsen, English, Hindi, Nagamese	Changki
AG	45	M	HSLC	Changki, Chungli, Mongsen, Nagamese	Changki
AM	34	F	PhD	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
NR	33	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
MD	40	F	HSSLC	Mongsen, Chungli, English, Nagamese	Mongsen
AY	42	F	HSSLC	Mongsen, Chungli, English, Nagamese	Mongsen
UN	52	M	HSLC	Mongsen, Chungli, English, Nagamese	Mongsen
UC	56	M	HSLC	Mongsen, Chungli, English, Nagamese	Mongsen
EH	45	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
NB	36	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AR	27	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
IM	32	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AK	27	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
IS	31	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AH	46	M	Grad.	Chungli, English, Nagamese	Chungli
IB	36	M	Grad.	Chungli, English, Hindi, Nagamese	Chungli
IN	26	M	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
AB	26	M	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli

Table B.4: Information of participants for perception test

Name	Age	Sex	Education	Language known	Mother tongue
SN	35	M	Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
AM	32	F	Post Grad.	Changki, Mongsen, Chungli, English, Hindi, Nagamese	Changki
AC	37	M	Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
AN	29	F	Grad.	Changki, Mongsen, Chungli, English, Hindi, Nagamese	Changki
ST	35	F	Grad.	Changki, Chungli, English, Hindi, Nagamese	Changki
IT	37	F	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
LN	32	F	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
SL	34	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
AM	33	F	Post Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen

CB	36	M	Grad.	Mongsen, Chungli, English, Hindi, Nagamese	Mongsen
NJ	34	F	PhD	Chungli, English, Hindi, Nagamese	Chungli
CL	50	F	HSLC	Chungli, English, Hindi, Nagamese	Chungli
AS	28	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
LN	28	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli
IN	27	F	Post Grad.	Chungli, English, Hindi, Nagamese	Chungli

List of Publications

Journals

1. **Moakala Tzudir**, Shikha Baghel, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Under-Resourced Dialect Identification in Ao using Source Information", The Journal of the Acoustical Society of America (JASA), vol. 152, no. 3, pp.1755-1766, 2022.

2. **Moakala Tzudir**, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Analysis and modeling of dialect information in Ao, a low resource language,"The Journal of the Acoustical Society of America (JASA), vol. 149, no. 5, pp.2976–2987, 2021.

Conferences

1. **Moakala Tzudir**, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Prosodic Information in Dialect Identification of a Tonal Language: The case of Ao," Proc. Interspeech 2022, 2238-2242.

2. **Moakala Tzudir**, Shikha Baghel, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Analyzing RMFCC Feature for Dialect Identification in Ao, an Under-Resourced Language," Proc. National Conference on Communications (NCC), pp. 308-313. IEEE, 2022.

3. **Moakala Tzudir**, Shikha Baghel, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Excitation Source Feature Based Dialect Identification in Ao—A Low Resource Language," Proc. Interspeech 2021, 1524-1528.

4. **Moakala Tzudir**, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Dialect Identification Using Tonal and Spectral Features in Two Dialects of Ao," in

Proc. Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages, 2018, pp. 137-141.

5. **Moakala Tzudir**, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Tonal Feature Based Dialect Discrimination in Two Dialects in Ao," in Proc. IEEE Region 10 Conference (TENCON) 2017, pp. 1795-1799.

Conferences (other than thesis work)

1. **Moakala Tzudir**, Mrinmoy Bhattacharjee, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Low-Resource Dialect Identification in Ao Using Noise Robust Mean Hilbert Envelope Coefficients," Proc. National Conference on Communications (NCC), pp. 256-261. IEEE, 2022.

2. Parismita Gogoi, **Moakala Tzudir**, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Automatic Tone Recognition of Ao Language," in Proc. 10^{th} International Conference on Speech Prosody 2020, pp. 1005–1008.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Bibliography

[1] Alexander Robertson Coupe. The Acoustic and Perceptual Features of Tone in the Tibeto-Burman Language Ao Naga. In *Proc. of the* 5^{th} *International Conference on Spoken Language Processing*, 1998.

[2] T Temsunungsang. Tonal correspondences in Ao languages of Nagaland. In 22^{nd} *Himalayan Languages Symposium.*, 2016.

[3] Fadi Biadsy, Julia Hirschberg, and Nizar Habash. Spoken Arabic dialect identification using phonotactic modeling. In *Proc. of the EACL Workshop on Computational Approaches to Semitic Languages*, pages 53–61, Stroudsburg, PA, USA, 2009.

[4] J. K. Chambers and P. Trudgill. *Dialectology*, volume 2^{nd} edition. Cambridge University press, 1998.

[5] A Etman and AA Louis. American dialect identification using phonotactic and prosodic features. In *SAI Intelligent Systems Conference (IntelliSys)*, pages 963–970. IEEE, 2015.

[6] *http://dcmsme.gov.in.*

[7] Walt Wolfram and Natalie Schilling. *American English: dialects and variation*, volume 25. John Wiley & Sons, 2015.

[8] A Etman and AA Louis Beex. Language and dialect identification: A survey. In *SAI Intelligent Systems Conference (IntelliSys)*, pages 220–231. IEEE, 2015.

[9] Yeshwant K Muthusamy, Etienne Barnard, and Ronald A Cole. Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41, 1994.

[10] Wanqiu Lin, Maulik Madhavi, Rohan Kumar Das, and Haizhou Li. Transformer-based Arabic dialect identification. In *International Conference on Asian Language Processing (IALP)*, pages 192–196. IEEE, 2020.

[11] Thimmaraja G Yadava and Haradagere Siddaramaiah Jayanna. A spoken query system for the agricultural commodity prices and weather information access in Kannada language. *International Journal of Speech Technology*, 20(3):635–644, 2017.

[12] Aditya Yadavalli, Ganesh Sai Mirishkar, and Anil Vuppala. Exploring the effect of dialect mismatched language models in Telugu automatic speech recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 292–301, 2022.

[13] Abhishek Dey, Abhash Deka, Siddika Imani, Barsha Deka, Rohit Sinha, SR Mahadeva Prasanna, Priyankoo Sarmah, K Samudravijaya, and SR Nirmala. Agroassam: A web based Assamese speech recognition application for retrieving agricultural commodity price and weather information. In *INTERSPEECH*, pages 3214–3215, 2018.

[14] George Abraham Grierson. *Linguistic survey of India*, volume 4. Office of the superintendent of government printing, India, 1906.

[15] Alexander Robertson Coupe et al. *A phonetic and phonological description of Ao: A Tibeto-Burman language of Nagaland, North-East India*. Pacific Linguistics, Research School of Pacific and Asian Studies, 2003.

[16] Directorate of census operation Nagaland. *District Census Handbook Mokokchung*. Nagaland, 2011.

[17] Mary M Clark. *The Ao Naga Grammar*. Assam Secretariat Printing Department, 1893.

[18] KS Gurubasave Gowda. *Ao-Naga phonetic reader*, volume 7. Central Institute of Indian Languages, 1972.

[19] K Gurubasave-Gowda. Ao grammar. *Mysore: Central Institute of Indian Languages*, 1975.

[20] E.W. Clark. *Ao-Naga dictionary*. Updated in 2013, 1911.

[21] Daniel Bruhn. The tonal classification of Chungli Ao verbs. *UC Berkeley PhonLab Annual Report*, 5(5), 2009.

[22] Alexander Robertson Coupe. *A Grammar of Mongsen Ao*, volume 39. Walter de Gruyter, 2007.

[23] T Temsunungsang. *The structure of Mongsen: Phonology and Morphology*. Hyderabad Central University MPhil thesis, 2003.

[24] Mohammad Abdel Qader Abu Shareah, Badri Abdulhakim DM Mudhsh, and Ayman Hamid AL-Takhayinh. An overview on dialectal variation. *International Journal of Scientific and Research Publications*, 2015.

[25] Yun Lei and John H. L. Hansen. Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:85–96, 2011.

[26] Wuei-He Tsai and Wen-Whei Chang. Chinese dialect identification using an acoustic-phonotactic model. In *Proc. of the* 6^{th} *European Conference on Speech Communication and Technology*, 1999.

[27] Bin Ma, Donglai Zhu, and Rong Tong. Chinese dialect identification using tone features based on pitch flux. In *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing*, volume 1, 2006.

[28] Dipanjan Nandi, Debadatta Pati, and K Sreenivasa Rao. Parametric representation of excitation source information for language identification. *Computer Speech & Language*, 41:88–115, 2017.

[29] *https://www.lexington.ro/tonal-vs-non-tonal-languages-chinese-vs-english/?lang=en.*

[30] Moira Yip. *Tone*. Cambridge University Press, 2002.

[31] Fadi Biadsy and Julia Hirschberg. Using prosody and phonotactics in Arabic dialect identification. In *Proc. of the* 10^{th} *Annual Conference of the International Speech Communication Association*, 2009.

[32] Jean-Luc Rouas. Automatic prosodic variations modeling for language and dialect discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1904–1911, 2007.

[33] Qian Zhang and John HL Hansen. Dialect recognition based on unsupervised bottleneck features. In *Proc. of the INTERSPEECH*, pages 2576–2580, 2017.

[34] Qian Zhang and John HL Hansen. Language/dialect recognition based on unsupervised deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5):873–882, 2018.

[35] Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. ADI17: A fine-grained Arabic dialect identification dataset. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248. IEEE, 2020.

[36] Marc A. Zissman, Terry P. Gleason, Deborah Rekart, and Beth L. Losiewicz. Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996.

[37] Pedro A. Torres-Carrasquillo, Terry P. Gleason, and Douglas A. Reynolds. Dialect identification using gaussian mixture models. In *Proc. of the Odyssey*, 2004.

[38] Rahul Chitturi and John HL Hansen. Multi-stream dialect classification using svm-gmm hybrid classifiers. In *Automatic Speech Recognition & Understanding, ASRU. IEEE Workshop on*, pages 431–436. IEEE, 2007.

[39] Rongqing Huang and John HL Hansen. Unsupervised discriminative training with application to dialect classification. *IEEE transactions on Audio, Speech, and Language processing*, 15(8):2444–2453, 2007.

[40] Nagaratna B Chittaragi and Shashidhar G Koolagudi. Automatic dialect identification system for Kannada language using single and ensemble svm algorithms. *Language Resources and Evaluation*, 54(2):553–585, 2020.

[41] Nagaratna B Chittaragi and Shashidhar G Koolagudi. Dialect identification using chroma-spectral shape features with ensemble technique. *Computer Speech & Language*, 70:101230, 2021.

[42] Rashmi Kethireddy, Sudarsana Reddy Kadiri, Paavo Alku, and Suryakanth V Gangashetty. Mel-weighted single frequency filtering spectrogram for dialect identification. *IEEE Access*, 8:174871–174879, 2020.

[43] Rashmi Kethireddy, Sudarsana Reddy Kadiri, and Suryakanth V Gangashetty. Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations. *The Journal of the Acoustical Society of America*, 151(2):1077–1092, 2022.

[44] Mohamed Ali. Character level convolutional neural network for german dialect identification. In *Proc. of the* 5^{th} *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 172–177, 2018.

[45] Shervin Malmasi and Marcos Zampieri. German dialect identification in interview transcriptions. In *Proc. of the* 4^{th} *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 164–169, 2017.

[46] Alina Maria Ciobanu, Shervin Malmasi, and Liviu P Dinu. German dialect identification using classifier ensembles. In *Proc. of the* 5^{th} *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 288–294, 2018.

[47] Jooyoung Lee, Kyungwha Kim, and Minhwa Chung. Korean dialect identification based on intonation modeling. In 24^{th} *Conference of O-COCOSDA*, pages 168–173. IEEE, 2021.

[48] K Sreenivasa Rao and Shashidhar G Koolagudi. Identification of Hindi dialects and emotions using spectral and prosodic features of speech. *IJSCI: International Journal of Systemics, Cybernetics and Informatics*, 9(4):24–33, 2011.

[49] Shweta Sinha, Aruna Jain, and SS Agrawal. Fusion of multi-stream speech features for dialect classification. *CSI transactions on ICT*, 2(4):243–252, 2015.

[50] Shyam S Agrawal, Aruna Jain, and Shweta Sinha. Analysis and modeling of acoustic information for automatic dialect classification. *International Journal of Speech Technology*, 19(3):593–609, 2016.

[51] Nagaratna B Chittaragi and Shashidhar G Koolagudi. Acoustic-phonetic feature based Kannada dialect identification from vowel sounds. *International Journal of Speech Technology*, 22(4):1099–1113, 2019.

[52] Wen-Whei Chang and Wuei-He Tsai. Chinese dialect identification using segmental and prosodic features. *The Journal of the Acoustical Society of America*, 108(4):1906–1913, 2000.

[53] Wuei-He Tsai and Wen-Whei Chang. Discriminative training of gaussian mixture bigram models with application to Chinese dialect identification. *Speech Communication*, 36(3):317–326, 2002.

[54] Gu Mingliang and Xia Yuguo. Chinese dialect identification using clustered support vector machine. In *International Conference on Neural Networks and Signal Processing*, pages 396–399. IEEE, 2008.

[55] Gu Mingliang, Xia Yuguo, and Yang Yiming. Semi-supervised learning based Chinese dialect identification. In *Proc. of the* 9^{th} *International Conference on Signal Processing*, pages 1608–1611. IEEE, 2008.

[56] Qiuxian Zhang, Yong Ma, Mingliang Gu, Yun Jin, Zhaodi Qi, Xinxin Ma, and Qing Zhou. End-to-end Chinese dialects identification in short utterances using cnn-bigru. In *Proc. of the* 8^{th} *Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 340–344. IEEE, 2019.

[57] Ravinder Singh and Anand Sharma. Identification system for different Punjabi dialects using random forest. *International Journal of Computer Sciences and Engineering*, 2018.

[58] Kanika Goyal, Amitoj Singh, and Virender Kadyan. A comparison of laryngeal effect in the dialects of Punjabi language. *Journal of Ambient Intelligence and Humanized Computing*, 13(5):2415–2428, 2022.

[59] Pham Ngoc Hung, Nguyen Thu Ha, Trinh Van Loan, Vu Xuan Thang, and Nguyen Dinh Chien. Vietnamese dialect identification on embedded system. *UTEHY Journal of Science and Technology*, 24:82–87, 2019.

[60] Mousmita Sarma and Kandarpa Kumar Sarma. Dialect identification from Assamese speech using prosodic features and a neuro fuzzy classifier. In *Proc. of the* 3^{rd} *International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 127–132. IEEE, 2016.

[61] Sofoklis Kakouros, Katri Hiovain, Martti Vainio, and Juraj Šimko. Dialect Identification of Spoken North Sámi Language Varieties Using Prosodic Features. *arXiv preprint arXiv:2003.10183*, 2020.

[62] Thangjam Clarinda Devi and Kabita Thaoroijam. Vowel-based Meeteilon dialect identification using a random forest classifier. *arXiv preprint arXiv:2107.13419*, 2021.

[63] Shikha Baghel, SR Mahadeva Prasanna, and Prithwijit Guha. Exploration of excitation source information for shouted and normal speech classification. *The Journal of the Acoustical Society of America*, 147(2):1250–1261, 2020.

[64] Shikha Baghel, SR Mahadeva Prasanna, and Prithwijit Guha. Excitation source feature for discriminating shouted and normal speech. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 167–171. IEEE, 2018.

[65] Yi Xu. Understanding tone from the perspective of production and perception. *Language and Linguistics*, 5(4):757–797, 2004.

[66] Seyed Omid Sadjadi and John HL Hansen. Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In *Proc. of the International conference on acoustics, speech and signal processing (ICASSP)*, pages 5448–5451. IEEE, 2011.

[67] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the Munich versatile and fast open-source audio feature extractor. In *Proc. of the* 18^{th} *ACM international conference on Multimedia*, pages 1459–1462, 2010.

[68] Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

[69] Moakala Tzudir, Priyankoo Sarmah, and SR Mahadeva Prasanna. Tonal feature based dialect discrimination in two dialects in Ao. In *Proc. of the Region 10 Conference, TENCON*, pages 1795–1799. IEEE, 2017.

[70] Moakala Tzudir, Priyankoo Sarmah, and S. R. Mahadeva Prasanna. Dialect identification using tonal and spectral features in two dialects of Ao. In *Proc. of the SLTU*, 2018.

[71] B Yegnanarayana and K Sri Rama Murty. Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):614–624, 2009.

[72] Phil Rose. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech communication*, 6(4):343–352, 1987.

[73] K Sri Rama Murty and B Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613, 2008.

[74] Vinay Kumar Mittal and B Yegnanarayana. Effect of glottal dynamics in the production of shouted speech. *The Journal of the Acoustical Society of America*, 133(5):3050–3061, 2013.

[75] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[76] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[77] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019.

[78] Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1):72–83, 1995.

[79] Nagaratna B Chittaragi, Asavari Limaye, NT Chandana, B Annappa, and Shashidhar G Koolagudi. Automatic text-independent Kannada dialect identification system. In *Information Systems Design and Intelligent Applications*, pages 79–87. Springer, 2019.

[80] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.

[81] Pedro A Torres-Carrasquillo, Elliot Singer, Mary A Kohler, Richard J Greene, Douglas A Reynolds, and John R Deller Jr. Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *Proc. of the 7^{th} international conference on spoken language processing*, 2002.

[82] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[83] Moakala Tzudir, Priyankoo Sarmah, and SR Mahadeva Prasanna. Analysis and modeling of dialect information in Ao, a low resource language. *The Journal of the Acoustical Society of America*, 149(5):2976–2987, 2021.

[84] Dipanjan Nandi, Debadatta Pati, and K Sreenivasa Rao. Implicit excitation source features for robust language identification. *International Journal of Speech Technology*, 18(3):459–477, 2015.

[85] Rohan Kumar Das and SR Mahadeva Prasanna. Exploring different attributes of source information for speaker verification with limited test data. *The Journal of the Acoustical Society of America*, 140(1):184–190, 2016.

[86] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[87] Roy D Patterson, KEN Robinson, John Holdsworth, Denis McKeown, C Zhang, and Michael Allerhand. Complex sounds and auditory images. In *Auditory physiology and perception*, pages 429–446. Elsevier, 1992.

[88] Sonia Gupta, Monika Agrawal, and Desh Deepak. Gammatonegram based triple classification of lung sounds using deep convolutional neural network with transfer learning. *Biomedical Signal Processing and Control*, 70:102947, 2021.

[89] Antonio Greco, Nicolai Petkov, Alessia Saggese, and Mario Vento. Aren: A deep learning approach for sound event recognition using a brain inspired representation. *IEEE Transactions on Information Forensics and Security*, 15:3610–3624, 2020.

[90] SR Mahadeva Prasanna, Cheedella S Gupta, and B Yegnanarayana. Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, 48(10):1243–1261, 2006.

[91] Xiaojia Zhao and DeLiang Wang. Analyzing noise robustness of MFCC and GFCC features in speaker identification. In *IEEE international conference on acoustics, speech and signal processing*, pages 7204–7208. IEEE, 2013.

[92] Brian R Glasberg and Brian CJ Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

[93] Sharu Goel, Sandeep Kumar Pandey, and Hanumant Singh Shekhawat. Analysis of emotional content in Indian political speeches. *arXiv preprint arXiv:2007.13325*, 2020.

[94] Rohan Kumar Das, Sarfaraz Jelil, and SR Mahadeva Prasanna. Significance of constraining text in limited data text-independent speaker verification. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2016.

[95] Steven F Sawyer. Analysis of variance: the fundamental concepts. *Journal of Manual & Manipulative Therapy*, 17(2):27E–38E, 2009.

[96] Wojtek Krzanowski. *Principles of multivariate analysis*, volume 23. OUP Oxford, 2000.

[97] Recommendation G.191 ITU-T. Software tools for speech and audio coding standardization, Int. Telecom. Union, Geneva, Switzerland, 2005.

[98] Recommendation G.191 ITU-T. ITU-T software tool library 2009 users manual," Int. Telecom. Union, Geneva, Switzerland, 2009.

[99] CM Vikram and SR Mahadeva Prasanna. Epoch extraction from telephone quality speech using single pole filter. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(3):624–636, 2017.

[100] E Vincent and DR Campbell. *Roomsimove*.

[101] Biswajit Dev Sarma, Abhishek Dey, Wendy Lalhminghlui, Parismita Gogoi, Priyankoo Sarmah, and S Prasanna. Robust Mizo digit recognition using data augmentation and tonal information. In *Proc. of the* 9^{th} *International Conference on Speech Prosody*, volume 2018, pages 621–625, 2018.

[102] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[103] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking. In *Proc. of the INTERSPEECH*, 2014.

[104] Florian Eyben, Felix Weninger, Martin Wöllmer, and B Shuller. Open-source media interpretation by large feature-space extraction. *TU Munchen, MMK*, 2016.