Analysis and Automatic Detection of Aspirated Fricative and Aspirated Nasals

Saswati Rabha

Analysis and Automatic Detection of Aspirated Fricative and Aspirated Nasals

A

thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

by

SASWATI RABHA

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

November 2022

Certificate

This is to certify that the thesis entitled "**Analysis and Automatic Detection of Aspirated Fricative and Aspirated Nasals** ", submitted by **Saswati Rabha**, Roll No. 156302013, 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 my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for the 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:

Guwahati.

Prof. S.R.M. Prasanna
Professor
Dept. of Electrical Engg.
Indian Institute of Technology Dharwad
Guwahati - 580 011, Karnataka, India.

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

To,

My Parents

(***Dr. Upen Rabha Hakacham***

and

Mrs. Manjula Rabha)

Acknowledgements

The thesis would not have been possible without the help and support of several people at various stages of my Ph.D. journey. First of all, I would like to thank God for giving me the strength and patience to complete my Ph.D. despite the numerous obstacles and having been infected twice by the virus during the COVID-19 pandemic.

I take this opportunity to express my heartfelt gratitude to my research supervisor, Prof. S.R.M. Prasanna. His continuous guidance in all aspects, constant motivation and support throughout the research journey were the driving forces that helped me complete my thesis. Besides academic support, his constant mental support when my entire family was infected with COVID-19 has helped me recover from some of the worst phases.

I would also like to thank my co-supervisor, Prof. Priyankoo Sarmah. His insightful feedback and valuable advice have helped me very much improve the quality of my thesis. I am also thankful to Dr. R. Sonkar and Dr. A. Ravindranath, the DPPC Secretaries of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their kind support as my administrative supervisor.

I am grateful to the esteemed members of my doctoral committee, Prof. S. Dandapat, Prof. Rohit Sinha, and Dr. Bidisha Som, for their constructive criticism, which helped me bring my work to the current form. I thank Prof. Samudravijaya for his valuable suggestions on my work. I would also like to thank all other faculty members of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their academic support. I am also very thankful to all the technical and non-technical staff members of the department and Jiban Barman Sir and Dr. L.N. Sharma Sir for their help when required.

My sincere thanks to my seniors (Dr. Biswajit Dev Sharma, Dr. Deepak K.T., Dr. Nagaraj Adiga, Dr. Rohan Das, Dr. Bidisha Sharma, Dr. Banriskhem K Khonglah, Dr. Sishir Kalita, Dr. C.M. Vikram, Dr. Protima Nomo Sudro, Dr. Akhilesh Dubey, Dr. Nagendra Kumar, Dr. Suman Deb, Dr. Ramesh K Bhukya, Dr. Abhishek, Dr. Subhasis Mandal, Dr. Himakshi Choudhury), past/present lab members (Dr. Vineeta, Dr. Alex, Abhishek, Moa, Mrinmoy, Shoubhik, Deepika, Brijmohan, Anik, Shikha, Samarjeet, Prabhakar, Mousumi, Salman, Sibasis, Debasish, Ato) and brothers (Dr. Ganji Sreeram, Dr. Tilendra Choudhury, Pradipta Sasmal, Debajit Sarma). I feel overwhelmed to get such wonderful people as my labmates who motivated me throughout my

research period with the constructive exchange of knowledge. I have been lucky to get my two close friends, Sandeep and Sukanya, who were my constant stress-busters and made my days in IIT beautiful. I would also like to thank Sarfaraz for always being there for me. Special thanks to Ali Jraisheh for being such a great friend and tolerating all my frustrations.

I thank Dr. Luke, Dr. Leena, Bidisha, Wendy from the Phonetics and Phonology Lab for sharing their knowledge. Sincere thanks to Viyazonuo Terhiija for helping me with the Angami database and Joyshree Chakraborty for her contribution to my thesis work. I am also thankful to Tabli, Sushanta, Nayanmoni, Nayan, Abhijit, Mohit, Nirab, Mrinal, Malek, and other friends for their help and support.

It should not be left unmentioned that I would not have enjoyed my IIT life without the aquatics club and the team of Bohagor Duwardolit. Swimming has helped me balance both physical and mental fitness throughout my Ph.D. life. I also thank Rajib Sir for giving me the opportunity to represent IIT Guwahati thrice in the INTER-IIT Aquatics meet. I want to thank all the IEEE Student Branch committee members, IIT Guwahati, for providing me a platform to explore and enrich my skill.

I am also very much indebted to my doctors, Dr. Pranjal Mahanta (orthopaedics), Dr. Paranjyoti Barman (neurosurgeon), Dr. P.P. Kalita and Dr. Hrishika (COVID doctors), doctor friends (Dr. Dalimi, Dr. Upama, Dr. Eva), my brother (Dr. Himajit) and sister-in-law (Dr. Purabi), for keeping me physically fit.

I take this opportunity to thank all the subjects who participated in this study. Thanks to Seung-ah Hong (Hankuk University of Foreign Studies, South Korea), Sangita, Utpal Da, Kamal Kumari aunt, the Rabha villagers for all the help during the database collection. I am very grateful for the contribution of Tankeswar Rabha, Chief of Rabha Hasong Autonomous Council, during my research period. Further, I would like to thank MHRD, Govt. of India, for providing the fellowship during my dual degree course (M.S.+Ph.D.).

Last but not least, my deepest gratitude goes to my family, Gourab, and his family. I am forever indebted to them for the love, support, understanding, sacrifice, constant blessings, and all the silent prayers for my success.

Saswati Rabha

Abstract

Unlike aspiration in stops, aspiration in non-stop consonants is quite rare. Most of the languages that have aspirated non-stop consonants are low-resource languages. Hence, data-driven, quantitative, and statistical analysis of their aspiration phenomena is fairly limited. From the literature review, it has been observed that there is still a need to explore a novel framework that will utilize the advantages of both linguistic and signal-processing knowledge-based approaches for acoustic-phonetic analysis and automatic classification of aspirated fricatives and aspirated nasals. To address these issues, we study two phonemes,/s/ and /s^h/ in a North-eastern language of India of Tibeto-Burman origin - Rabha, where contrast exists between aspirated and unaspirated counterparts. Also, for the study of aspirated nasals (/m/- /m^h/, /n/-/n^h/ and /ŋ/-/$ŋ^h$), we choose the Angami language, another North-eastern language of India of Tibeto-Burman origin. Both languages are low-resourced and are characterized by these unique sounds.

Initially, a detailed discussion about the development of a speech database for the thesis work and the procedure of database preparation is provided. In order to study the aspirated and unaspirated fricatives and nasals, the database was created for the Rabha and Angami as there is no standard built-in database available in the public domain for those languages. Data was collected from the Korean speakers as no open-source Korean database was available that was relevant to our area of interest. While the reviewed literature presents Korean fricatives as two-way contrast fricatives: aspirated and unaspirated, and the presence of the two-way contrast of the Angami nasals has been reported, no study has been attempted for the phonological fricative contrast in Rabha. Hence, this perception experiment would give an insight into the existence or non-existence of phonemic pairs of Rabha fricatives, where aspiration may serve as a cue. The discrimination test and identification test of the perception experiment

confirm that the Rabha fricatives are phonemically distinguished in terms of aspiration. Further, production analysis on the recorded speech stimuli of the fricatives and nasals was conducted.

The subsequent work unearths the details of aspiration characteristics from linguistic studies and further utilizes acoustic-phonetic information to extract source features using signal processing methods. A set of acoustic features is proposed to detect aspiration in fricatives and nasals automatically. Acoustic features, such as vocal tract constriction (VTC), normalized autocorrelation peak strength (NAPS), the strength of excitation (SoE), and variance of successive epoch intervals (VSEI) are used to detect aspiration in fricatives and nasals. Results show that VTC, NAPS, and SoE can detect aspiration in the nasals, whereas SoE and VSEI can detect aspiration in fricatives.

Further, to capture the distinctive aspiration associated with the fricatives, the source features, along-with durational features, were used to model a two-class classification system. The extracted features are combined to improve the classification of fricatives based on aspiration. The acoustic parameters are fed to support vector machine (SVM), Gaussian mixture model (GMM), and Random Forest (RF) classifiers. The integrated features fed to the SVM-GMM-RF hybrid model outperform baseline MFCCs for classifying fricatives in Rabha and Korean, irrespective of the language.

Another work investigates the nature of peak ERB_N trajectories to analyze the spectral dynamics across the time course of fricatives and nasals. The work is motivated by the fact that differences in the articulation of the two-way contrast non-stop consonants lead to differences in the contours of peak ERB_N in the aspirated and unaspirated variants. Such differences help to propose an efficient approach for analyzing and automatically classifying the aspirated fricatives and aspirated nasals from their unaspirated counterparts. The experimental investigation has shown that combining the static and dynamic acoustic parameters improved the classification accuracy compared to the baseline MFCC features.

Finally, we propose a framework that can utilize different acoustic-phonetic knowledge at

various stages to classify aspirated and unaspirated non-stop consonants automatically. Also, possible applications of the proposed methods are discussed by demonstrating the details of an aspiration detection-based phone recognition system on the collected corpus. The proposed method improves the performance of an automatic phoneme recognizer by reducing the confusion between aspirated and unaspirated counterparts.

Keywords: Fricatives, Nasals, Aspiration, under-resourced language, speech recognition, speech analysis, perception, low-resourced languages, Rabha, Angami, Korean.

Contents

List of Figures

List of Tables

List of Acronyms

AM	acoustic model
AP	acoustic-phonetic
ASR	automatic speech recognition
EMST	Electro-Medical and Speech Technology
ERB	equivalent rectangular bandwidth
ERB_N	equivalent rectangular bandwidth number
FBANK	filter-bank
GMM	Gaussian mixture model
HMI	human-machine interaction
HMM	hidden Markov model
HTK	hidden Markov model toolkit
IITG	Indian Institute of Technology Guwahati
LM	language model
LPC	linear prediction coefficient
MFCC	mel-frequency cepstral coefficient
NAPS	normalized autocorrelation peak strength
OOV	out-of-vocabulary
PRS	phone recognition system
SOE	strength of excitation
SVM	support vector machine
RF	Random Forest
VAD	voice activity detection
VSEI	variance of successive epoch intervals

VTC	vocal tract constriction
WER	word error rate

1

Introduction

Contents

"One of the most difficult aspects of performing research in speech processing is its interdisciplinary nature and the tendency of most researchers to apply a monolithic approach to individual problems [3]." Computer technology, artificial intelligence, digital signal processing, pattern recognition, acoustics, linguistics, and cognitive science are some of the disciplines involved with the study of speech. Phonetics and speech signal processing are the most widely used discipline for studying speech signals. Phonetics is a branch of linguistics that studies the physical properties of speech sounds and broadly deals with two aspects of human speech: production—the ways humans make sounds—and perception—the way speech is understood. On the other hand, the relevant information is extracted from the speech signal efficiently, using different machine learning techniques through signal processing. When speech sound is studied from a linguistic point of view, more importance is given to the phonetic theories of the acoustics and articulation of the sounds and little importance to the well-defined, automatic procedures for detecting speech sound units. On the contrary, for speech signal processing study, there is a tendency to focus more on explicit automatic feature extraction and pattern recognition approach rather than the speech-specific knowledge. Therefore this approach is relatively insensitive to speech intelligibility and misses out on the acoustic-phonetic information [3].

Speech is produced and perceived by human beings as a sequence of events. The events during production are articulatory in nature due to the physical movements of various articulators of the speech production apparatus. The signatures of these articulatory events are reflected in the speech signal and are perceived as acoustic events by a listener. Hence acoustic-phonetic (AP) segmentation of speech by detecting these acoustic events in the speech signal seems to be the most natural way of analyzing the phonetic content in the speech signal. Deriving the acoustic-phonetic information embedded in the speech signal with the help of a linguistic approach and further using that knowledge in automatizing a detection task is, therefore, an important task in speech signal processing.

Acoustics of speech sounds is studied by phoneticians and most often it is achieved by using speech analysis tools, such as, PRAAT [4]. While such tools provide inbuilt routines for several types of speech analysis based on signal processing methods, the possibility of customization of such routines is restricted. This limits the ability to analyze and model speech properties that

are unique or infrequent. In such cases, custom-tailored signal processing methods can be used to capture features of rare and unique speech sounds. Additionally, machine learning algorithms can help in the modeling and classification of speech sounds, more effectively in well-resourced languages and with some efforts in under-resourced languages.

As mentioned earlier, the modeling of unique sounds in under-resourced languages requires effort not only from the machine learning approach but also from acoustic phonetics. Hence, the work in this thesis firstly involves in-depth acoustic analysis of the unique sounds. The knowledge from the acoustic analysis forms the basis for feature selection for identification and classification of the sounds. Lastly, machine-learning (ML) based methods are used for automatic detection of phonetic features, such as, aspiration, frication, etc. The final part of this thesis demonstrates the application of phonetics-based approach and signal processing-based approach in automatic speech recognition system. In the next section, an overview of the unique sounds that exist in the under-resourced languages of the world is provided.

1.1 Unique sounds in low-resource languages

Many languages may contain unique and interesting sounds, apart from the commonly occurring speech sounds, which represents characteristic regional features. The under-documented low-resource languages have several unique characteristics, such as tone system, presence of unique sounds, etc. It is important to sample as many diverse sounds and languages as possible to gain a better understanding of variations across and within languages and the limits and typology of speech sounds, and sound combinations [5].

Speech sounds are broadly divided into two categories: vowels and consonants (stops, nasals, fricatives, affricates). Consonants are further categorized based on place of articulation and manner of articulation. The place of articulation is the place where the obstruction in the vocal tract takes place (bilabial, dental, alveolar, palatal, velar). The manner of articulation is the configuration and interaction of the articulators which result in various types of airstream in the production of the consonants. For instance, the airstream is completely blocked in the production of a consonant, such as, /b/ or /p/. On the other hand, airstream is released in a restricted manner, resulting in noise-like turbulances in the production of fricatives, such as, /s/ or /f/.

This thesis primarily deals with the speech characteristics of consonants that have a two way contrast based on aspiration. Aspiration is a typo-logically rare, under-studied, but important phonemic feature in several Indian languages, making it an interesting study subject. From an articulatory point of view, aspiration is considered a puff of air or"turbulent noise" between the articulation of a consonant and the start of voicing, which result in a whispery, breathy, and voiceless speech [6]. When there is an aspiration, the resonances of the vocal tract are excited by steady airflow from the wide-open glottis [3]. 'Aspiration' refers to diffuse turbulent noise before the modal vowel, and this aspiration can be either voiced, or voiceless [7]. It has been observed that among the consonants, aspiration contrast in stops is the most common in the Indian languages and is widely studied. However, the aspiration contrast is also observed in some of the unique sounds, such as aspirated fricatives and aspirated nasals. When the fricatives and the nasals are followed by noise-like aspiration characteristics during articulation and production of the sound, such speech sounds are said to be aspirated fricatives and aspirated nasals.

Aspirated nasals and aspirated fricatives are typo-logically rare in world's languages, and the literature to characterize them is very limited [8–10]. While the total number of living languages in the world cannot be known precisely, a total of 7,117 living languages is listed in Ethnologue, the most extensive catalog of the world's languages [11]. Among them, 2186 languages are reported in the PHOIBLE database, and [12] out of those, only 1% (26) confirmed the existence of aspirated fricative in their consonant inventories. Reports of occurrence and studies concerned with the acoustic characterization of aspirated fricatives are scarce as they occur in some of the under-documented languages such as Bodo, Rabha, Burmese, Sgaw Karen, Pumi, Qiádōng Hmongic, Central Mazahua, Cross River Mbembe, Barbarino and Embera-Catio [12–16]. While the aspirated fricatives in Korean have been well-documented, limited studies have been done in Rabha fricatives, as the Rabha language itself is a low-resource language and is not yet documented in the PHOIBLE database. Therefore, we will study the acoustic-phonetic information embedded in these sounds and use this information for building automatic methods for the detection of these unique sounds.

Another rare and unique sound, aspirated nasal, is reported in 0.2% of the total languages mentioned in the PHOIBLE database. Among the Indian languages, the presence of aspirated nasals has been reported in Marathi, an Indo-Aryan language [17] and Angami, a Tibeto-Burman

language of the North-East India [15]. It is to be noted that aspirated nasals are not the same as a nasal murmur. While some define the nasals with aspiration as aspirated nasals [9, 15] as well as breathy nasals [17, 18], in a few studies, the nasals have been reported as two types - voiced and voiceless nasals [19–21]. Angami nasals have been described as aspirated with positive voice onset time [15,22], however in another study, the Angami nasals are described as voiceless nasals [20]. The current state-of-the-art is relatively limited in the analysis of Angami nasals, and in the application of machine learning algorithms for their phonemic classification.

Therefore, two under-documented languages from North-East India, viz., Rabha, and Angami, are chosen for study in this thesis. Both languages are of Tibeto-Burman origin, low-resourced, and under-developed, characterized by these unique sounds. We attempt to provide a broader picture of the unique sounds in these languages by utilizing knowledge from both linguistic and signal processing fields. Extensive use of phonetics-based approach and signal processing-based approach has been used to understand the unique sounds of the languages as mentioned above. To sum up, this thesis aims to find an efficient technique for analysis and automatic classification of the unique sounds: aspirated fricatives and aspirated nasals from their unaspirated counterparts using machine learning.

1.2 Motivation for the thesis work

Aspiration results in contrastive phonemes in several languages. The contrastive phonemes occur as minimal pairs, i.e., they are two similar-sounding words such as a phoneme, toneme, or chroneme that differ in only one phonological element and have distinct meanings in a particular language. They are used to demonstrate that two phones are two separate phonemes in the language. Hence, in a language that has a phonemic contrast between a pair of aspirated and unaspirated speech sounds, automatic identification of the aspiration feature is crucial in its phoneme recognition system. Let us see an example. The phonemic contrast in the nasals /m^{h}a/ and /ma/ is a pair of words in Angami language that differs in aspiration feature but constitutes two different word meaning ‘something to work’ and ‘people’, respectively. Similarly /n^{h}o/ and /no/ are another minimal pair meaning ‘plastic cover’ and ‘child’ respectively. In Korean language, the two-way contrast fricatives /s^{h}an/ and /san/ means ‘mountain’ and ‘cheap’ respectively. While [p] and [p^{h}] are not phonemically

contrastive in English, they contrast phonemically in Hindi where /pal/ means 'time', /p^{h}al/ would mean 'fruit'. This distinction of aspirated-unaspirated sounds is important in the phone recognition application of these languages. Such a facility could also be useful for the detection of pronunciation errors for non-native learners who may not have been exposed to the aspiration-based distinction in their native tongue [23].

When two-way contrast consonants occur in a particular language and are wrongly detected, it poses many challenges in automatic speech recognition and affects the intelligibility of the speech sounds. If the aspirated stop consonant is recognised as the unaspirated variant /pal/ instead of /p^{h}al/ in Hindi language, *"Aam bohut mit^{h}a* ***p^{h}al*** *he"* would translate to

"Mango is a very sweet ***time****"* (instead of)

"Mango is a very sweet ***fruit****"* (right).

Such a "microscopic" approach to speech analysis not only helps in automatic speech recognition but also has numerous applications as highlighted by Bozena Kostek [24]: *"pre-lexical processing in spoken-word recognition, automatic evaluation of pronunciation [25], differences between speakers' native regional accents [26], or the evaluation of speech articulatory disorders [27] as it allows a more in-depth speech analysis."*

While phonemically contrasting aspirated and unaspirated stops are fairly common in languages, only a handful of languages have phonemic contrasts between aspirated and unaspirated nasals and fricatives and the literature to characterize them is very limited [8]. These non-stop consonants typically have a two-way aspiration contrast. However, its phonetic implementation is less clear-cut, isolating acoustic cues that reliably define the difference, particularly the typo-logically rare "non-stop aspirates," has proven challenging [7]. The relatively less-studied, rare, and unique sounds aspirated fricatives, and aspirated nasals are generally problematic to define and pose difficulties. The challenges here are linked to a lack of understanding of the articulation of the aspiration in aspirated fricatives and aspirated nasals and limited resource constraints for building a powerful language model that represents the vocabulary, grammar, and typical usage of the low-resource language.

Except for Korean, most languages with phonemic non-stop aspiration are low-resource languages that lack relevant speech data and detailed acoustic analysis. Therefore, a need arises to

develop a database for these low-resource languages and understand the acoustic-phonetics and aerodynamics of the typo-logically rare and unique aspirated fricatives and aspirated nasals. In India, the presence of aspiration in voiceless nasals and fricatives is reported in the Tibeto-Burman languages, which are primarily spoken in north-east India [15]. The aspirated fricatives in the Tibeto-Burman languages are postulated to have emerged from the **tsh* in Proto-Tibeto-Burman (PTB) [1]. In this thesis, Rabha and Angami languages, both low-resource Tibeto-Burman languages spoken in northeast India, are chosen to investigate the characteristics of the unique aspirated fricatives and aspirated nasals. The presence of aspirated fricatives is also attested in Bodo [16], another Tibeto-Burman language, contiguous to Rabha.

While there have been many acoustic studies of the fricative contrast across the Korean language, there has been almost no work investigating which cues listeners use to distinguish the Rabha fricatives and Angami nasals from their unaspirated variants in perception. In the existing literature, Rabha and Angami have been extensively studied from ethnographical, historical, and language documentation perspectives, but there is a dearth of phonetic and signal processing studies. Whether the fricatives of Rabha have a breathy interval (aspiration) like other major Tibeto-Burman languages requires in-depth investigation based on quantitative phonetic data. Here, we present the first instrumental phonetic description of the controversial two-way contrast (alveolar unaspirated, aspirated) fricatives of Rabha. Since the phonemic status of unvoiced aspirates of Rabha fricatives is controversial, our main focus is to investigate which acoustic and articulatory correlates distinguish the unvoiced aspirates from plain unvoiced consonants while also taking into account automated approaches for feature extraction. Thus, the current study is the first detailed investigation of the two-way fricative contrasts and their automatic detection and classification in an under-documented Rabha language. An investigation into the acoustic properties of Rabha offers an empirical contribution to the understanding of fricatives cross-linguistically. We also present the first combined study of signal processing and phonetic description of the two-way contrast Angami nasals.

Thus, the primary goal of this work is to provide an acoustic-phonetic description of the aspirated fricatives and aspirated nasals in Rabha and Angami, respectively. Next, there is a requirement to develop automatic methods to detect unique sounds in under-documented languages. There should be a feature extraction mechanism for the automated approach that focuses on the discriminatory

cues for the unique sounds having two-way contrast in aspiration. Different machine learning algorithms need to be explored for detecting acoustic-phonetic events like aspiration in aspirated fricatives and aspirated nasals. This leads to the secondary goal of this work, i.e., to consider the acoustic-phonetic information and develop an automated method for the detection of these unique sounds in Rabha and Angami. A signal processing approach is attempted for the acoustically analyzed speech utterances to accomplish the secondary goal. In an automatic approach such as automatic speech recognition (ASR), Mel-frequency cepstral coefficients are the most commonly used feature. However, this feature does not pay attention to discriminatory sounds. On the other hand, acoustic-phonetic approaches emphasize these discriminatory sounds and help categorize them. Therefore, automatic signal processing and machine learning methods are used to detect these unique sounds out of the different sound units. We will study the acoustic phonetics of these sounds and use this information to build automatic methods for detecting these unique sounds. In this work,acoustic-phonetic features, motivated by an understanding of the production of the aspirated non-stop consonants, are evaluated for the classification of fricatives and nasals along the aspiration dimension. Several new acoustic measures are proposed for the reliable detection of the aspiration contrast in fricatives and nasals.

1.3 Challenges

Different speech sounds can be characterized by a set of spectral and temporal properties that depend on the acoustic-phonetic features of the sound. When these acoustic-phonetic features are added to conventionally used features, it enhances the performance of a basic recognition procedure. The first step in the acoustic-phonetic approach to speech recognition is the speech analysis system. Here, the characteristics of the time-varying speech signal are represented by a set of features specific to that particular phonetic unit in a spoken language. Hence, the next step is the feature detection and extraction system. It involves algorithms to characterize different phonetic units such as nasality, frication, formant locations, voiced-unvoiced classification, and ratios of high and low-frequency energy [3]. Followed by this step is the segmentation and labeling phase, where the speech signal is segmented into discrete regions and labeled according to the acoustic properties of the segmented region.

This acoustic-phonetic approach has challenges of its own. The greatest challenge in the speech recognition process for an under-documented or low-resource language is the non-availability of the database in the public domain. For a low-resource language, Rabha, spoken in the north-eastern part of India, the only mode of learning the language is verbal as there is no formal educational training in the Rabha language. Also, co-habiting with Assamese and Garo language speakers influences the native Rabha language speakers, thus affecting the originality of the language. Even after the collection and development of the database, the segmentation and labeling phase of the speech recognizer is the most difficult one to carry out reliably. No well-defined, automatic procedure exists for tuning the method on labeled speech in a consistent manner which poses another challenge [3].

Moreover, aspiration in non-stop consonants is more difficult to perceive than stops. Sometimes, human perception is strong enough to distinguish between aspirated and unaspirated counterparts. However, due to failure to capture certain characteristics of aspirated fricatives, there might be a mismatch between the acoustic and perceptual characterization of aspirated fricative [28]. Unlike longer VOT of aspirated stops, shorter VOT of aspirated fricative makes it harder to perceive the distinction between aspirated and unaspirated counterparts [29]. As seen from the diagrams in Figure 2.1 the frication to aspiration transition is not always distinct compared to stops. It isn't easy to distinguish between frication and aspiration phenomena by seeing only time domain or frequency domain characteristics. Also, the source mechanism, i.e., turbulence noise, remaining the same for both fricative and aspiration, might pose a signal processing challenge in the automatic classification of aspirated fricative. So, before we go for automatic classification, the spectro-temporal behavior of the underlying aspiration with aspirated fricatives is to be appropriately analyzed by the signal processing method. This will help in segmentation of the boundaries of fricative and aspiration, which, in turn, will help in classifying aspirated fricative from its unaspirated counterparts, based on the presence/absence of aspiration segment. No signal processing tool has been addressed so far in the existing literature to automatically mark the onset of aspiration in the case of aspirated fricative. Hence, all the aforementioned issues are addressed in this thesis.

1.4 Objectives and scope of thesis

The work addresses a few research questions: It seems that aspirated and unaspirated fricatives are in free variation in Rabha, unlike in Korean. Then, why do we need to detect aspiration in the first place? And how do we even know whether the detection is accurate or not? The logic seems a bit circular - the fricative must first be identified as aspirated or not to know whether aspiration was correctly detected, but how is aspiration identified without a detection algorithm? If it is easy enough to do visually, then why do we need to detect it? This situation seems completely different from Korean, in which there is a phonological fricative contrast, for which aspiration may serve as a cue. This work tries to answer all these questions and also explains the contribution of the Rabha data here with perceptual analysis to verify that aspiration has any linguistic consequence. Another research question that arises is how similar or different the aspiration characteristics in the aspirated nasals are compared to that of the aspirated fricatives. Will the features that work for automatic detection of aspirated fricatives also work for the aspirated nasals? This thesis tackles all these research questions.

This thesis studies the production of human speech sounds by acoustic modeling and signal analysis. It concentrates on less studied, rare, and unique sounds among the world's low-resources languages. The unique sounds considered in this work include fricatives and nasals that have two-way contrast based on aspiration. The languages chosen to study these sounds are Rabha and Angami, two under-documented languages. The current state-of-the-art is relatively limited in Rabha fricatives, and there is no open-source database for the Rabha language. So, the first attempt of this work is to create a database of the Rabha language by recording and transcribing the speech corpus and then proceeded by perceptual and acoustic analysis. Speech processing techniques are fairly limited for the classification of the Angami nasals and Rabha fricatives. The main focus of this thesis is to develop signal processing methods for the analysis and detection of different acoustic-phonetic events in the aspirated fricatives and aspirated nasals. The signal processing method involves the technique of extracting relevant informants using static and dynamic properties of the speech signal in an efficient, robust manner. Therefore, in this thesis, an attempt has been made to understand the properties of the aspirated fricatives and aspirated nasals, extract the acoustic features that capture the distinct information of the two-way contrast in these non-stop consonants,

and develop a simpler non-stop consonants classifier that would automatically detect and classify the aspirated fricatives and aspirated nasals from their unaspirated variants.

The objectives of the current study can be broadly classified as:

- To study and analyse the spectral characteristics of rare and unique sounds- aspirated fricatives (s-s^h) and aspirated nasals (m-m^h, n-n^h, ŋ-$ŋ^h$)
- In order to study aspirated fricatives in a low-resource language, Rabha, an attempt will be made to create the Rabha speech corpus containing scripted read speech data recorded from the native speakers of Rabha language. This attempt would fill the void of the absence of any standard open-source database available for the Rabha language.
- The study would require extensive knowledge of interdisciplinary nature. Therefore, an attempt is made to understand the production, perception, and acoustic-phonetic characterization of the aspirated non-stop consonant sounds by incorporating the signal processing techniques as well as the methodology used in linguistics
- To find an efficient approach for extraction of features specific to the target phonetic units and automatic classification of the aspirated fricatives and aspirated nasals from their unaspirated counterparts using machine learning.

1.5 Organization of the thesis

The rest of the thesis is organized as follows. A brief literature review on the aspiration in stop consonants, fricatives, and nasals is presented in Chapter 2. In Chapter 3, the collection of speech databases for low-resource languages, Rabha and Angami, and the Korean database and the development of the corresponding lexical resources are described. Perception of the fricatives and the acoustic cues for characterization of the aspirated and unaspirated fricatives and the nasals have also been presented in this chapter. The state-of-the-art techniques employed for training various systems using source features to classify aspirated-unaspirated data are described in detail in Chapter 4. The temporal features and the source features are explored in Chapter 5. Chapter 6 explores the spectral analysis of the fricatives and nasals varying over time. The details of the

baseline ASR system development on the collected corpus and the obtained experimental results by combining auditory and articulatory-based features are reported in Chapter 7. Finally, the thesis is summarized, and the possible future directions of the work are discussed in Chapter 8.

2

Aspiration in speech sounds: A review

Contents

2.1 Introduction

In phonetics, aspiration is the strong burst of breath that accompanies either the release or, in the case of preaspiration, the closure of some obstruents [3]. Aspiration is found in English, German, Hindi, Urdu, Bangla, etc. [12]. In the International Phonetic Alphabet (IPA), aspirated consonants are written using the symbols for voiceless consonants followed by the aspiration modifier letter [^{h}], a superscript form of the symbol for the voiceless glottal fricative /h/ [30]. For instance, /p/ represents the voiceless bilabial stop, and /p^{h}/ represents the aspirated bilabial stop. Many Indo-Aryan languages have aspirated stops. Sanskrit, Hindi, Bengali, Marathi, and Gujarati have a four-way distinction in stops: voiceless unaspirated, voiceless aspirated, voiced unaspirated, and breathy-voiced or voiced aspirated, such as /p ph b bh/. For e.g /pal/ /phal/ means "time" and "fruit" respectively in Hindi. In Assamese, /bat/ means"way" while /bhat/ means "rice".

Different literature suggests different definitions of aspiration, as mentioned in the following. From an articulatory phonetics point of view, aspiration is defined as the unvoiced excitation inside the voice box/glottis [31] or quick projection of air released after voiceless stops /p/, /t/, /k/ respectively when they are found in stressed syllables and before vowel sounds [32]. Acoustically, aspiration is defined as an interval of aperiodic noise of glottal origin [29] which resembles a short [h] fricative before voicing of vowel [3].

Aspiration is an important phonemic feature that is found commonly in stops and rarely in fricatives and nasals [12, 22, 33–35]. However, there exists a small body of study that explored the acoustic characteristics and methods of detection and classification of aspiration in fricatives and nasals occurring in some of the under-documented languages [14, 15, 29, 36]. This chapter reviews the literature related to the acoustic-phonetic analysis of speech sounds associated with aspiration and the approaches that use the knowledge of these analyses. The following sections discuss the articulatory and acoustic characteristics of aspiration in consonants with three articulation manners: stop, fricative, and nasal. As the relevant literature that focuses on aspiration detection mechanisms in the case of fricatives and nasals are very few, attempts for in-depth analysis and review of fricatives and nasals have been conducted in the context of existing literature available for stops. Due to the limited availability of literature, we explore in what other contexts aspiration in other consonants has been studied and see if those methods would help in the study of aspiration in fricatives and

nasals.

2.2 Articulatory characteristics of aspiration

The literature related to the study of articulation characteristics of aspiration in different consonants, i.e., stops, nasals and fricatives, are presented in the following subsections.

2.2.1 Aspiration in Stop Consonants

Aspiration occurs in stop consonants as a period of voicelessness after stop release. In case of English, aspiration is associated only with /k/, /t/ and /p/ consonants [8, 37]. Nevertheless, aspiration in stops in English occurs when they are in stressed syllables and the onset position. At the moment of stop release, the open glottis allows the breath-stream to flow freely into the upper vocal tract without generating phonation [3]. Voiceless aspiration occurs when the vocal folds remain open after a consonant is released [38]. When the aspirated interval is included, aspirated stops are significantly longer than the unaspirated ones. Hence, the increased stop duration is compensated by reduced vowel duration [39]. The increased duration is significant as aspiration needs to be of sufficient length for a stop to be correctly perceived as an aspirated one [40]. The spectrograms of an aspirated and unaspirated stop in Figure 2.1 (a) and Figure 2.1 (d) show that the aspirated one has longer aspiration between the release of the stop and following vowel onset.

Results from a previous study of Korean stops have reported that firstly, lenis, fortis, and aspirated stops are systematically differentiated by the voice quality of the following vowel. Secondly, these stops are also differentiated by aerodynamic mechanisms. The aspirated and fortis stops are similar in supralaryngeal articulation but employ a different relation between intraoral pressure and flow [41]. It was observed that in Sgaw Karen, aspirated stops feature an interval of aspiration between the release of the stop (signaled by the short stop burst in the spectrogram) and the beginning of the following vowel (signaled by the onset of periodic waves in the spectrogram) [29]. Another study confirmed that the intra-oral pressure and intra-oral airflow are the highest for aspirated stops [41].

2.2.2 Aspiration in Fricatives

Although most aspirated obstruents in the world's languages are stops and affricates, aspirated fricatives such as [sʰ], [fʰ] or [ɕʰ] have been documented in Korean, in a few Tibeto-Burman languages, in the Siouan language Ofo and in some Oto-Manguean languages [14]. Except for Korean, most of the languages that have phonemic non-stop aspiration are low-resource languages that lack relevant speech data and detailed acoustic analysis. Aspirated fricatives are reported in only 1% of the total languages mentioned in PHOIBLE database [12]. According to Craioveanu (2013) [13], additional evidence of phonemic aspirated fricatives exists for Burmese and other languages such as Sgaw Karen, Pumi, Tibetan, Southern Subanen, etc. Some languages, such as Choni Tibetan, have up to four contrastive aspirated fricatives [sʰ] [ɕʰ], [ʂʰ] and [xʰ][3]. Korean is one of the languages with phonological contrast between /sʰ/ and /s/. Some describe the contrasts as /sʰ/ vs. /s/ [42,43] while others use /s/-/sʰ/ [41,44], /s_F/-/s_{NF}/ [34] or /s/-/s'/ [45,46]. These differences in notations to represent the /sʰ/ vs. /s/ contrasts may be due to the discrepancies in description of the fricative contrasts in IPA. The two contrasting views are, while some are of the opinion that Korean has (1) lenis and fortis denti-alveolar fricatives, the others are of the view that Korean has a (2) aspirated vs. fortis contrast. In Korean, /sʰ/ has been classified as fortis fricative [-s.g, +tense] whereas /sʰ/ or /s/ has been specified as lenis fricative [+s.g, -tense] [45, 47]. The lenis or non-fortis fricative is aspirated with broader and shorter oral constriction and shorter pharyngeal width as compared to /sʰ/ [45, 48]. The non-fortis Korean fricative has been analyzed in various ways in the literature — as aspirated by some [49–51], as lenis by others [35, 41, 42, 52], and as both aspirated and lenis by others still [44, 53]. In this study, based on aspiration, the authors have used the notation /s/ (unaspirated category) and /sʰ/(aspirated category) for the between-fricative differences in Rabha and Korean.

As seen in Figure 2.1, panel (e), an aspirated fricative has frication followed by aspiration. While the following vowel is distinguishable owing to its periodicity, the preceding aspiration and frication parts are difficult to tell apart from each other as both are characterized by aperiodic noise resulting from turbulent airflow. However, in Sgaw Karen, it is claimed that the aperiodicity in aspirated fricatives is of two types. While the coronal frication phase is characterized by high amplitude and higher resonant frequencies, the aspiration phase of glottal origin is indicated by noticeably

lower amplitudes, and resonant frequencies [29]. In the production of a fricative sound, turbulent noise is produced at a constriction in the vocal tract. While frication noise is produced above the glottis, aspiration noise is produced at the glottis. In languages that have the contrast of aspiration in fricatives, plain unvoiced fricatives are considered as [-spread glottis] as the duration of glottis spreading is less than that of aspirated fricatives [14]. Korean is the only well-studied language with contrasting aspirated, and unaspirated voiceless alveolar fricatives [40, page 80]. A preliminary hypothesis on the articulation of aspiration contrasts is provided [14], based on three overlapping gestures involved in the articulation: constriction, upstream/downstream pressure adjustment, and beginning of voicing of the vowel. In both aspirated and unaspirated fricatives, frication starts with a combination of increased intra-oral pressure and oral constriction. In unaspirated fricatives, upstream pressure drops simultaneously with the constriction release [14]. However, there is a delay in releasing the constriction in aspirated fricatives, and there is a drop in the upstream pressure. The constriction is released while upstream pressure from the lungs is maintained, resulting in aspiration. It is to be noted that the sound source of both frication and aspiration noise is turbulence noise. Turbulent noise is produced as a result of constriction in the vocal tract. Aspiration noise is produced at the glottis. Frication noise is produced above the glottis.

2.2.3 Aspiration in Nasals

Aspirated nasals, as a distinct phonemic category, is very rare in languages. The PHOIBLE database reports only 6 languages with aspirated nasal sounds [12]. Among the languages that have voicing contrasts in nasals, Burmese, Kham Tibetan, and Xumi (all from the Tibeto-Burman language family) are relatively well-studied [21, 54, 55]. The categorization of the nasals based on aspiration contrast has been controversial. The Angami nasals are sometimes categorized as voiceless and sometimes as aspirated [15, 20, 22, 42, 49, 50]. It is reported that Burmese has voiced nasals, /m, n, ɲ, ŋ/, contrasting with /m̥, n̥, ɲ̥, ŋ̥/, in terms of voicing while voiceless aspirated nasals /m/ [m^h] and /n/ [n^h] are reported in case of Xumi. While spectrographic analyses on Burmese voiceless nasals showed that the voiceless nasals consist of a portion of voiceless nasal friction and a voiced nasal portion [54], however, the Burmese voiceless nasals were reported to be voiceless in the initial portion and voiced towards their terminal portions [19]. The energy of the voiceless nasal friction

portion is much lower than that of the voiced nasal portion. Additionally, the duration of the voiced portion is found to be shorter than that of the nasal friction [54]. Based on these observations, it is claimed that there are two types of voiceless nasals in the world, one the Burmese type and the other the Angami type (a Tibeto-Burman language of north-east India), which has continuous nasal airflow with no voicing throughout the production of the voiceless nasal [19].

Among the Indian languages, the presence of aspirated nasals has been reported in Marathi, an Indo-Aryan language [17] and Angami, a Tibeto-Burman language of the north-east India [15]. Instrumental studies on Angami voiceless nasals report the absence of voicing during their production. Aerodynamic studies have reported the beginning of oral airflow a little over halfway through the voiceless portion [15, 22]. Two types of voiceless nasals reported in the languages of the world are voiceless unaspirated nasals and voiceless aspirated nasals [56]. The first type is the preaspirated nasals which contain a voiceless portion with nasal and oral airflow, followed by a voiced portion with only nasal airflow. The second type is the voiceless nasals with continuous nasal airflow throughout the duration of the nasal. However, towards the end of such nasals, oral airflow is added and characterized by partial voicing. The Angami nasals have been reported as aspirated [15] which categorically belongs to the second type of nasals. In a recent work, the two-way contrast in Angami nasals has been categorized as voiceless and voiced nasals, in which oral aspiration is associated with the voiceless nasals [20]. Nasometric studies have been conducted, and the amount of voicing, nasalance, and duration were measured. The results showed that the voiceless nasals produced in isolation were mostly voiceless, with oral aspiration appearing around the middle of the articulation. When produced in sentence frames, an additional voiced nasal portion was noticed at the onset of the voiceless nasal. In terms of nasality, determined by nasalance scores, the effect of the place of articulation was observed. It was also noticed that the onset of oral aspiration resulted in reduced nasalance. The voiceless nasals were consistently longer than their voiced counterparts. Generally, when produced in sentence frames, all the nasals were longer, regardless of voicing [20]. In the case of voiceless nasals, Angami shows distinctive patterns when there are produced in isolation and sentence frames. While the ones produced in isolation contain primarily voiceless nasal and oral aspiration, in sentence frames, voiceless aspiration is preceded by voiced nasal portions [20].

Aspirated nasals are reported in another language, Sukuma of Eastern Bantu region [9] and

the four aspirated nasals occur in bilabial, alveolar, palatal, and velar place of articulation. While initially Sukuma nasals were categorized as voiceless nasals followed by [h], one of the speakers classifies them as voiced nasals with "strong," "very strong," or "very noticeable" aspiration. The phonetic observations with the help of nasal airflow, oral airflow, and intra-oral air pressure measurement revealed that aspiration portion is usually voiced and voiced nasal portion is 2-3 times longer than prenasalized voiceless aspirated stops [9].

The articulation of aspiration in nasal consonants differs from that of oral consonants. Nasals are produced with the air-stream stopped in the oral cavity passed through the nasal cavity. They are produced with glottal excitation, and the vocal tract is constricted at some point along the oral passageway [57]. Rate of airflow is observed to be faster after aspirated nasals(/n^h/) compared to unaspirated nasals(/n/) [15].

2.2.4 Summary

Detection of aspiration is relatively easy when it occurs after a stop consonant, as the burst associated stops provide a reasonably clear boundary that separates the aspiration from the stop. One study reported that in the case of aspirated stops, the period of voicelessness following the release of a stop serves as a distinct acoustic feature. This is accompanied by a distinctive burst that provides sufficient acoustic evidence for identifying the boundary between the aspiration and the burst [58]. However, in the case of the non-stop aspirated consonants, acoustic features are not very distinctive. The properties of aspiration associated with stops are significantly different from those associated with fricatives and nasals. Unlike stops, an overlapping gesture of aspiration and frication or aspiration and nasal is observed in the fricatives and nasals. Distinguishing between an aspirated and an unaspirated nasal is difficult due to the presence of strong nasal resonance in the higher frequency bands for both types of nasals [59]. In the initial work of this thesis, it was observed that, due to the overlapping gestures of aspiration and the nasals, determining the exact boundary between aspiration and nasal is a formidable task. Figure 2.1 shows that the aspiration transition is not distinct due to the short interval when compared to stops. In the case of an aspirated nasal, lack of burst makes detecting the boundary between the nasal portion and the aspiration part difficult. Moreover, the source for both frication and aspiration is turbulent

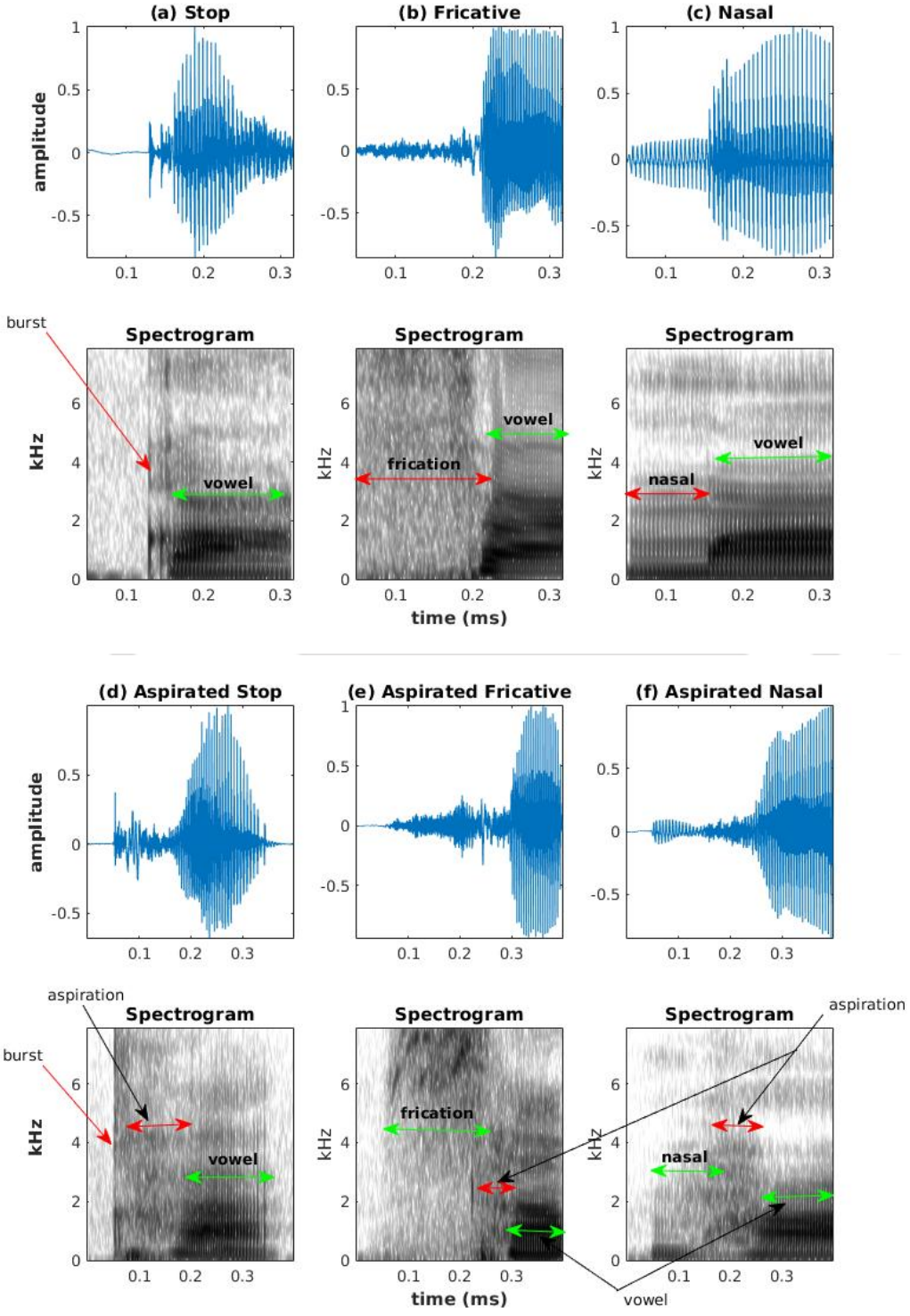

Figure 2.1: Waveform and spectrogram of (a) /ka/ (b) /sa/ (c) /ma/ (d) /k^{h}a/ (e) /s^{h}a/ (f) /m^{h}a/

noise and contains similar spectral characteristics. Therefore, identifying the boundary between the aspirated portion and the frication portion of an aspirated fricative also poses a challenge. In the case of all types of fricatives, determining the boundary between the noise and voicing is challenging as there is no definite boundary between the two [40]. The overlapping area between noise and voicing is also recognized as a distinct transient area in the literature [60]. As seen in Figure 2.1, while the boundaries of the aspiration region in the aspirated stop consonants can be determined relatively easily, the same is not the case for aspirated nasals and aspirated fricatives. Another interesting phenomenon that was observed was that the presence of aspiration after the frication /s^{h}/ is found in non-high vowel environments, especially before /a/ and barely present in high vowel contexts [46, 51, 61–64].

Considering the preceding discussion, we examine the Fourier spectra of aspirated and unaspirated fricatives for evidence of contrasting spectral cues. As seen in Figure 2.2, we notice differences in the spectral quality of the aspiration and frication parts of the aspirated fricatives. Based on the generalized observation of the aspirated fricative database used in the study, we conclude that aspiration has a higher energy concentration in the low-frequency regions (< 500 Hz). In comparison, frication has a higher energy concentration in the high-frequency regions (> 4000 Hz). Similar observations have also been reported in [29].

In order to see the potential differences in the spectrum for nasal and aspiration regions, the Fourier spectra of aspirated nasals are inspected as seen in Figure 2.2. It was observed from the spectra that the aspiration region contains components with higher energy in the $500 - 1200$ Hz range. On the other hand, the nasal portion of the phoneme has a continuously falling spectrum throughout the frequency range of $0 - 4000$ Hz.

2.3 Acoustic correlates of aspiration

2.3.1 Stops

Various acoustic-phonetic cues have been suggested for the characterization of aspiration in stop consonants. Lisker and Abramson (1964) [37] suggest that VOT is the most familiar model for classifying aspirated stop consonants. VOT was also used in another study for the task of detecting aspirated stops as an aspirated-unaspirated distinction task[[33], [23, 65]]. Some authors

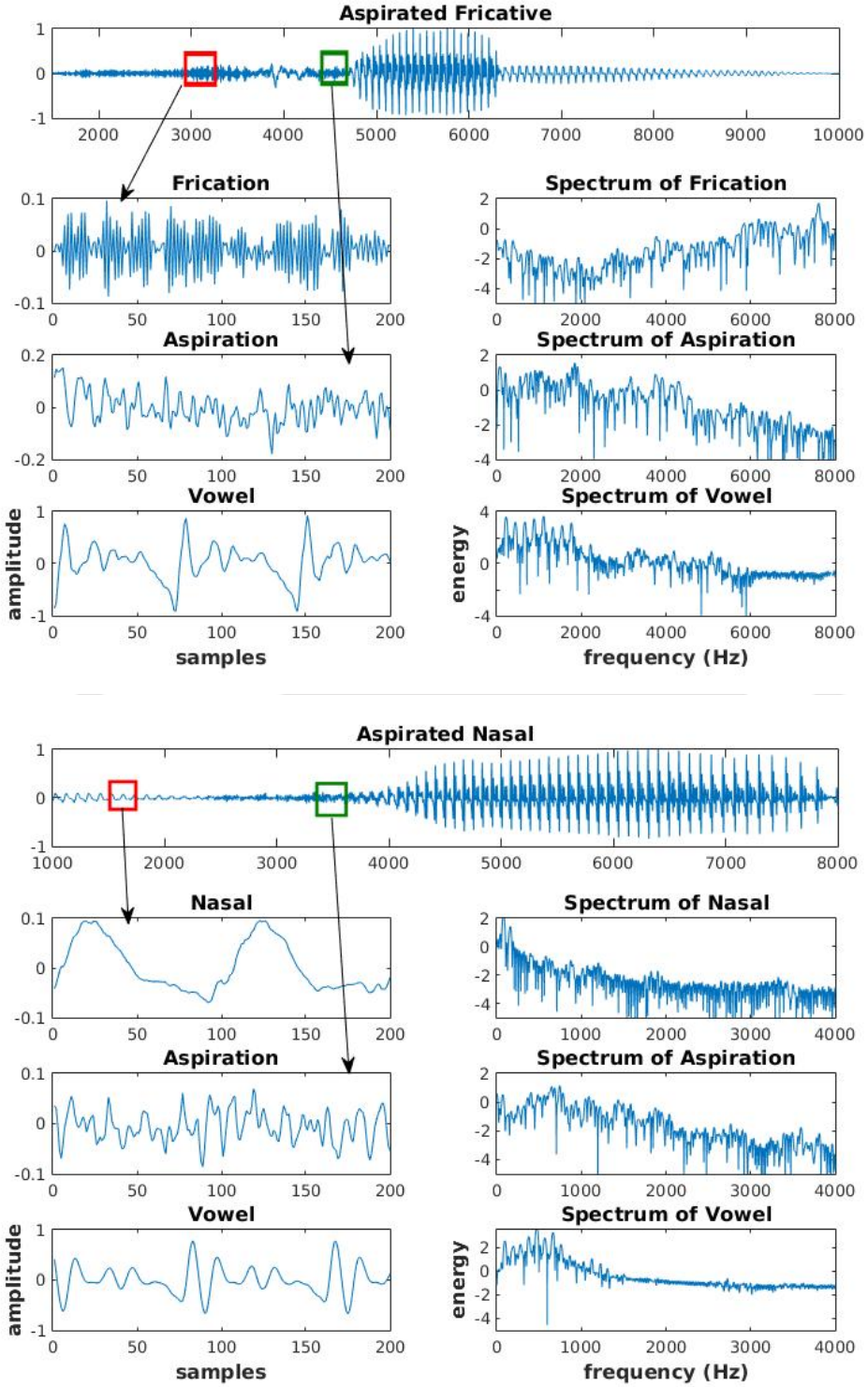

Figure 2.2: Spectral energy-based classification for aspirated fricative and aspirated nasal respectively

have pointed out that this method is not able to distinguish the plain voiced and voiced aspirated stops [66]. In a study for Korean stops, it was shown that that relative burst energy distinguishes the aspirated stops from the unaspirated ones [41].

Analysis of the closure duration of segments, measuring the beginning of the burst or its release, F0 lowering, structure of the lower harmonics, spectral tilt etc. were other measures used as cues for aspirated stops of Nepalese language [67]. Several acoustic measures such as lag-VOT, H1–H2, A1–A3, SNR, F1F3-sync, Low-band-slope, B3-band energy,pre-onset energy ratio, are proposed for the reliable detection of the aspiration contrast in unvoiced and voiced plosives along the aspiration dimension in Hindi and Marathi language [23, 33, 65, 68]. The features capture the release characteristics and the distinctive glottal pulse shape and aspiration noise in the vowel region following the voiced and unvoiced aspirated plosives.

2.3.2 Fricatives

For Korean aspirated fricatives, the behavior in phonetic processes and its phonetic realizations are generally believed to be similar to stops in the aspirated category [49–51]. In some of the previous studies, [35, 41, 52], acoustic measures such as frication duration, the centroid of the fricative noise, fundamental frequency, H1-H2, and H1-F2 were used to understand the aerodynamic characteristics of the two-way distinction between the voiceless fricatives /s/ and /s^{h}/. In another study of Korean fricatives, it has been stated that aspiration is closely connected with gradual intensity buildup, breathy voice onset, and high F1 onset [34, 44, 61]. This point is significant because it implies that these cues may be inevitable consequences of the glottal configuration required for aspiration. Another measure, aspiration duration, was found to be the most significant cue with a longer period of aspiration creating the perception of /s^{h}/ [51,62]. Also, the perception experiments suggested that the native listeners were less accurate at identifying /s^{h}/ and /s/ in high vowel contexts [51, 61, 62] where the aspiration does not differentiate the categories very well (e.g. /s^{h}i/-/si/ and /s^{h}u/-/su/) [62, 63].

While F0 and frication duration did not contribute anything to Korean fricative contrast, aspiration duration, spectral mean, and H1-H2 all contributed to the /s^{h}/-/s/ contrast [41]. Another work reports that /s^{h}/ has higher value of aspiration duration and H1-H2, whereas /s/ has a higher

value for the spectral mean [62]. Perception study by native Korean speakers from Seoul, Daegu, Jeju, and native Mandarin and Japanese suggest that cues to the Korean fricative contrast are stronger in low vowel contexts than in high vowel contexts [63].

Apart from the above-mentioned measures, skewness, location of spectral peaks, spectral peak frequency have been studied to investigate aspirated fricatives in some languages such as Korean and Bodo [16,41]. In Bodo language, measures such as vowel duration, aspiration duration, frication duration, average vowel intensity, F0 at three locations, F1 onset, center of gravity, Spectral tilt(H1-H2) have been used to characterize aspirated fricative [16]. However, disparities in acoustic and perceptual similarity results suggest the requirement of other acoustic dimensions for understanding the characteristics of aspirated fricatives [28]. Apart from these features, acoustic cues such as spectral peak location, spectral moments, amplitude, F2 at transition have been explored in a few fricative studies [69].

2.3.3 Nasals

For the study of the two-way contrast in nasals, articulatory phonetics has been attempted in previous studies. Acoustic and aerodynamic measurements such as the total duration of the target phonemes, the duration of the voiced period during the target phonemes, mean nasal and oral flow have been conducted on the nasals of the languages such as Xumi, Burmese and Tibetan [21]. Apart from the acoustic cues such as the amount of voicing, duration [56], and nasalance [20] have been used for distinction of two-way contrast of nasals, little study has been attempted for the acoustic correlates of aspirated nasals. As such, we will be looking into acoustic correlates for the linguistic feature nasal to distinguish nasals as a class from other classes of sounds. Pruthi [70] focused on the extraction of parameters (such as energy onsets/offsets, spectral peak frequency, energy ratio, envelope variance measure) related to the nasal murmur, while [71] used formant based parameters to detect nasal consonants in sonorant regions. Fujimura [72,73] identified four properties that characterize the spectra of nasal murmurs: (1) the existence of a very low first formant at about 300 Hz, (2) the existence of zeros in the spectrum, (3) relatively high damping factors of the formants, and (4) high density of formants. Features such as peak amplitudes in five frequency bands were measured: $A1n(0\text{–}788Hz), A2n(788Hz\text{–}2kHz), A3n(2\text{–}3kHz), A4n(3\text{–}4kHz), and A5n(4\text{–}5kHz)$. The sum of

the differences in amplitudes between $A1n$ and other amplitudes, and the frequency of the lowest peak $F1n$ were used as the murmur parameters [74, 75]. Chen [74] did a more exhaustive work on the acoustic correlates of nasality. Relative energy change in the frequency bands is used to categorize nasals from the non-nasals [76]. However, these parameters for nasals are studied to distinguish nasals from other classes of sounds, and none of the languages report the presence of aspirated nasals. The relevant information from these studies is utilized for developing acoustic parameters that would characterize aspirated nasals. An acoustic-phonetic analysis of the aspirated nasals would facilitate the extraction of parameters for automatic detection and classification of aspirated-unaspirated nasals.

2.3.4 Summary

Studies on phonemic aspiration in the world's languages have mainly been confined to unvoiced consonants, and voice onset time has been the chief distinctive feature explored [33, 37, 52]. For a given language, identification of aspirated and unaspirated consonants is important for the applications such as identification of native and nonnative speakers, analysis of phonological processes in children [77], learning the cultural evolution of any language [78], improving the performance of automatic speech recognition system (ASR) [79] and so on. This phenomenon of aspiration and unaspiration of sounds can be identified by capturing some pronunciation-specific cues [80]. While such studies have been done for aspirated stop consonants, to the author's best knowledge, no previous attempt has been made to automate the classification approach of fricatives and nasals based on aspiration, specifically for the low resource languages.

An acoustic analysis of plain and breathy voiced obstruents and sonorants in Marathi was conducted in order to test whether breathy phonation in sonorants is associated with the acoustic correlates of breathiness found for vowels and obstruents [17]. However, more importance was given to the distinction based on the place of articulation rather than the manner of articulation. The acoustic differences between plain and breathy consonants are diminished in sonorant contexts: [b] differs from [b^h] more than [m] differs from [m^h] [17]. Compared to aspirated-unaspirated pairs of stops, the lesser acoustic difference between the nasal pairs and fricative pairs does not rule out the fact that the nasals have two-way contrast. The two-way contrast, if ignored, can pose a

challenge in phoneme classification in automatic speech recognition. That is, if in a given language, a phoneme were swapped with another phoneme, it would change the meaning of the word. For example, the English words kid and kit end with two distinct phonemes, and swapping one for the other would change the word's meaning. However, the difference between the p sounds in pun (p^h, with aspiration) and spun (p, no aspiration) never affects the meaning of a word in English, so they are phones and not phonemes. By contrast, swapping the same two sounds in Hindi or Urdu can change one word into another: /p^hal/ means 'fruit', and /pal/ means 'moment'. Similarly, in languages like Korean, where aspirated fricatives exist, the meaning changes based on aspiration. For e.g /sa/ means 'to buy' whereas /s^ha/ 'to wrap', /su/ implies 'number' while /s^hu/ 'cook' [51]. Classification of aspirated consonants from unaspirated counterparts is important because of their phonemic significance. Therefore the acoustic-phonetic study of nasals and fricatives requires separate investigation from linguistic and signal processing approaches.

As studied by Chang [34,44,61], acoustic features of /s^h/ and /s/ are produced with (i) different durations (the fortis fricative being longer than the non-fortis fricative), (ii) different aspiration intervals (the nonfortis fricative's being longer than the fortis fricative's), (iii) different F1 trajectories (F1 after the fortis fricative starting lower than after the non-fortis fricative), (iv) different patterns of intensity buildup (intensity increasing more rapidly after the fortis fricative than after the non-fortis fricative), and (v) different voice onset qualities (a more breathy quality after the non-fortis fricative than after the non-fortis fricative, as indicated by a more steeply declining spectral tilt). On the other hand, f0 onset, average vowel intensity, and vowel length do not appear to be distinguishing factors [41]. Salient acoustic properties of aspiration in fricatives are considered to be vowel duration, Centre of gravity (COG), F1 onset, spectral tilt (H1-H2), F0 [28,34]. Spectral mean, which reflects the size of the front cavity of a fricative [81], was found to be a specific cue for the differences in the two fricatives for Korean language [62,63]. It was found that the spectral mean of an aspirated fricative (associated with more linguo-palatal contact) is higher than that of an unaspirated one (associated with smaller front cavity) [41,82,83]. Studies related to acoustic correlates of aspirated nasals are relatively few. Amount of voicing and duration are the two acoustic cues that have distinguished the aspirated nasals from the unaspirated ones in the previous research studies.

2.4 Automated method for the classification of aspirated-unaspirated category

2.4.1 Stops

Due to the lack of resources for the automatic detection of non-stop consonants based on aspiration contrast, we explore the signal processing methods that have been used for automatic classification through the context of stops. The task of detecting aspirated stops as an aspirated-unaspirated distinction task was investigated as a 2-class problem [[33], [23,65]]. Various acoustic features such as VOT, Spectral features (HI-H2, HI-A3, and AI-A3), Synchronization index, Sub-band spectral power, and sub-band slope, Signal to noise ratio (SNR) were investigated for aspiration detection in voiced and unvoiced stops of Marathi [23]. The performance was evaluated for various features such as landmark-based acoustic features [23, 33, 65, 84], joint spectro-temporal features [85, 86], Gabor features [87] in a classification of place of articulation in unvoiced stops. These spectral and spectro-temporal features when compared with the MFCC-HMM baseline system for the recognition of aspirated and unaspirated stops were found to be more robust than the MFCC baseline ASR system [88].

It was observed that for Hindi and Marathi stops, acoustic-phonetic features, motivated by an understanding of the production of the aspirated plosives, were evaluated for the classification of plosives along the aspiration dimension [23, 33, 65]]. The acoustic-phonetic features are compared with more standard automatic speech recognition features in a baseline MFCC-HMM system for classification based on aspiration. The acoustic-phonetic features are shown to perform well in the two-way classification task and also appear robust to cross-lingual training and testing, as demonstrated by a classification experiment using Marathi speech-trained acoustic models on native Hindi utterances.

2.4.2 Fricatives

While several works have reported the acoustic differences between aspirated and unaspirated fricatives [36, 41, 44, 48, 89], only a few have attempted to automatically detect aspiration in aspirated fricatives. In one approach [48], the best classification accuracy of 79.74% was reported for the Korean alveolar fricative class in bi-gram modeling using HTK toolkit and F0 cues. However,

the work does not mention the confusion between the two types of fricatives, namely, aspirated and unaspirated. In [90, 91], the authors investigated the spectro-temporal variation, productions of English /s/ and /ʃ/ and of Japanese /s/ and /ɕ/ in word-initial, prevocalic position were elicited from adult native speakers. The spectral dynamics of these productions were analyzed in terms of a psychoacoustic measure of peak frequency: "peak ERB_N number." to determine how peak frequency varied temporally within each fricative. Three analyses compared (1) the English sibilants to each other, (2) the Japanese sibilants to each other, and (3) English /s/ to Japanese /s/. The results indicated that, in both English and Japanese, the sibilant fricatives differ acoustically in terms of both static (i.e., overall level) and dynamic (i.e., shape) aspects of the peak $\mathrm{ERB_N}$ number trajectories. Furthermore, English /s/ and Japanese /s/ exhibited language-specific differences in the shape, but not overall level, of peak $\mathrm{ERB_N}$ number trajectories. While this work has demonstrated the potential of categorizing the sibilant fricatives, the categorization of aspiration contrast in fricatives was not attempted.

Significant research interest has been focused on detecting the place of articulation of fricatives [81, 92, 93]. However, studies directed towards automatically detecting the aspirated fricatives from the speech are minimal. Most of the studies address the articulatory and acoustic phonetics of aspirated fricatives, but the acoustic parameters are not extracted automatically. Due to the lack of studies on an automated method for detecting aspirated-unaspirated fricatives, we focus on automatically detecting fricatives as a broad class from continuous speech. Spectral and temporal evidence like dominant resonant frequency, epoch strength, and numerator of group delay at zero-frequency, band-energy ratio [94–96] have been used as cues to detect the unvoiced fricatives from continuous speech. Vydana and Vuppala(2016) [97] have used S-transform-based time-frequency representation for detecting fricatives from continuous speech. The method localizes the high-frequency events with time, and it was observed that S-transform-based time-frequency representation performs better than STFT in detecting broad fricative classification. This method proposed by Vydana detects the broad fricative class (sibilant fricatives, non-sibilant fricatives, affricates) from the continuous speech of the TIMIT database. However, for a fricative with comparatively weak high-frequency energy (aspirated fricatives), the approaches that rely on capturing the frequency distribution information (S-transform and STFT based approaches) for detecting fricatives perform poorly. In

another study, Support vector machines (SVMs) and Recurrent neural networks (RNNs) trained with 39-dimensional MFCCs with appropriate framing and frequency shift parameters are employed for detecting fricative broad class [95, 98–101]. The previous methods have been able to detect the fricative reliably well for the fricative broad-class classification, but none addresses the issue of extracting the acoustic correlates of aspirated fricatives automatically.

2.4.3 Nasals

Previous studies [38, 72, 73, 102, 103] have provided us with a wealth of information about the acoustic manifestations of the nasal consonant, but few have addressed the issue of extracting the acoustic correlates of nasality and aspirated nasals automatically. Glass (1984) and Glass and Zue (1986) [104, 105] did some experiments on the automatic extraction of acoustic parameters for the distinction between nasals and a class of impostors consisting of liquids, glides, voice bars, and weak voiced fricatives. They used the following parameters: (1) difference in average energy between the consonant and adjacent vowel, (2) percentage of time a resonance was centered between 200 and 400 Hz, (3) average amplitude of the resonance in the region between 200 and 400 Hz relative to the total energy in the consonant, (4) ratio of energy between 0 and 350 Hz to energy between 350 and 1000 Hz, and (5) change in low-frequency energy throughout the consonant. While a 74% average detection rate was obtained for nasalized vowels in one automated approach [104], in another method, the acoustic parameters were used for automatic detection of nasal manner, which resulted in classification accuracies of 89.53%, 95.80% and 87.82% for prevocalic, postvocalic, and intervocalic sonorant consonants respectively on the TIMIT database [70]. However, the parameters for nasalization do not address the issue of extracting the acoustic correlates of aspirated nasals automatically. As such, there is a requirement to develop acoustic parameters that characterize aspirated nasals specifically.

2.4.4 Summary

With respect to aspiration in consonants, most of the works presented in the literature are focused on the automatic classification of stops based on place of articulation and manner of articulation. The fricatives classification method is focused on the place of articulation rather than distinguishing the aspirated fricatives from the unaspirated ones. The same is observed in the case of nasals too.

Detection of aspirated fricatives and nasals is rarely described in the literature. These works include analysis of stop consonants along with fricatives and affricates.

It was also clear that the VOT could play a significant role in aspiration detection as observed in the case of stops; however, it is not expected to be able to perform the task alone. Other features are needed to resolve the significant overlap that exists between aspirated and unaspirated counterparts, especially for fricatives. The previous study has shown that acoustic measures perform better than MFCC feature-based systems in differentiating aspirated-unaspirated classes of stops. Therefore, the same is expected in the case of aspirated fricatives, including that of the nasals. MFCC feature is used for the representation of transition information. But MFCCs may not be suitable for the analysis of transition region as they do not capture temporal dynamics explicitly [84,86]. Apart from MFCCs, joint spectro-temporal cues were proposed to analyze stop consonants [84,106]. Computation of spectro-temporal features does not require explicit knowledge of formant frequencies. While the exploration of spectro-temporal features has been explored in distinguishing aspirated-unaspirated pairs of stops, such studies have not been done in fricatives and nasals, which motivates us to look in that direction. This work attempts to develop a knowledge-based speech recognition system that uses landmarks to search for distinctive features. In the speech signal, landmarks identify times when the acoustic manifestations of the linguistically motivated distinctive features are most salient [107].

The features (MFCCs or acoustic parameters) considered for development and evaluation of a speaker-independent automatic system for the robust detection of aspiration in stop consonants in the previous approaches are more reliable for vocal tract system characteristics of the speech signal, but features representing excitation source characteristics of the speech signal are not much explored. Also, most of the experiments were conducted to investigate the contrast between the two voiceless sibilant fricatives of Korean, while no previous attempt has been made to detect the aspiration in Rabha aspirated fricatives in Rabha and Angami aspirated nasals. The excitation source (i.e., vocal folds) and the vocal tract system interact significantly with each other during the production of speech. Hence cues based on excitation source information and spectral features might help provide additional support for speech analysis and representation.

Based on the review of previous work and our experiments, we have narrowed down on the

following : (a) energy information, (b)vocal tract information (c) spectral information. Significant research interest has been focused on detecting the place of articulation of fricatives and nasals [70,97,108]. However, studies directed toward automatically detecting the fricative regions and nasal regions from the speech are minimal, in particular when the fricatives and nasals contrast based on aspiration . The information that exists in the literature is neither sufficient nor consistent enough to be integrated into the automatic classification of aspirated fricatives. Therefore, the thesis focuses on developing relevant acoustic parameters (APs) that can be reliably and automatically extracted and used to automatically classify two-way contrast in fricatives and nasals based on aspiration.

2.5 Issues in Existing Approaches and Plan of study

The results from the previous studies prompt us to look more closely into periodicity, vocal fold configuration, and vocal tract constriction patterns to detect aspiration in the non-stop consonants. Except for Korean, most languages with phonemic non-stop aspiration are low-resource languages that lack relevant speech data and detailed acoustic analysis. Considering this, in this work, we explore the possibility of using spectro-temporal features in these rare non-stop aspirated sounds to distinguish them from their unaspirated counterparts.

As reviewed in the previous section, several studies have examined several acoustic and aerodynamic parameters that might differentiate the three-way contrast in Korean stops and two-way contrast in the stops of the languages of Indo-Aryan origin [41]. Most previous studies have studied the aerodynamic characteristics of two-way contrast in stops compared to those of non-stops (fricatives and nasals). Nevertheless, as far as we know, there is no single work documented in the literature that addresses all the acoustic-phonetic parameters and challenges of automation with a relatively large subject pool in aspirated fricatives and aspirated nasals.

The study is expected to exploit different features like source, system, and suprasegmental information to extract aspirated fricatives and aspirated nasals and mark their different segment boundaries. Automatic classification of aspirated fricative will be helpful as a part of a segmentation and phoneme categorization system and in addressing transcription issues in the ASR system. This thesis work attempts to tackle an important research problem: lack of research for acoustic-phonetic features specific to the unique sounds: aspirated fricatives and aspirated nasals and its

demonstration in automatic methods of classification of fricatives and nasals from their unaspirated counterparts.

One more purpose for the study of these unique sounds is their rare complex phenomenon and their presence almost exclusively in less-studied languages among the world's languages [14]. Another merit in examining the Rabha and Angami non-stop consonants lies in documenting a low-resource language and preparing and developing a database for the low-resource language. Even for a widely-studied language like Korean, it is worthy of an intense study investigating the limits of human speech capabilities as the Korean consonant system is unique among the world's languages. The complexity of the phonetic properties of the three-way contrastive Korean stops implies that speech production is far more complicated than we can predict from simplified phonological representations or even the more sophisticated general speech production models that are currently available [41]. Studying a low-resource language is even more challenging due to the low availability of the database. Due to the non-teaching of Rabha language in the elementary school level education, multi-dialectal effect, and the influence of other language speakers in the neighborhood region, it is not easy to find pure native speakers of Rabha in the city. With the help of local Rabha dialect experts, we were able to find a few native speakers of Rabha in the villages of Assam and Meghalaya. All of them still spoke Rabha, uninfluenced by other languages.

Though the Korean fricatives are widely studied, the availability of open-source databases is relatively low. Most of the languages that have aspirated non-stop consonants are low-resource languages. Hence, data-driven, quantitative, and statistical analysis of their aspiration phenomena is relatively limited. Rabha and Angami languages are considered in this study, as previous studies have confirmed the existence of aspiration contrast in fricatives and nasals. To understand the phonetic properties underlying the two-way contrastive fricatives in Rabha and Korean, we created our speech database by collecting a body of data from a relatively large number of speakers and examining a variety of phonetic parameters for both of the collected corpora in a similar process. Without such controls, we cannot be assured about the generalizability of the data. In the present study, we collected speech data from native speakers of Rabha, Korean, and Angami and examined a variety of phonetic parameters simultaneously. Although the size of the sample in the present study may not be large enough to generalize to the entire population, it is expected that the present study

will serve as an improved reference for future research related to aspirated fricatives and aspirated nasals in low-resource languages by providing ample data from various phonetic parameters.

As there are limited studies on aspirated nasals, in this thesis work, we try to investigate the characteristics of the aspirated nasal by studying its features and comparing them with the characteristics of the aspirated fricatives. The goal is to check the similarity or dissimilarity of the acoustic features of aspiration in the rarely studied aspirated nasals with that of the aspiration in the previously studied aspirated fricatives [16, 44, 52]. There is a two-way contrast in the nasals of Angami as well as in the case of Rabha fricatives - aspirated and unaspirated. The focus of this study is the phonetic analysis of the Angami phonemes /m/, /n/, /ŋ/ and /m^h/, /n^h/, /$ŋ^h$/ having different place of articulation (i.e labial, alveolar and velar origin respectively). Since the alveolar fricative in Korean too has two-way contrast - /s/ and /s^h/ i.e., unaspirated and aspirated counterparts, hence an acoustic comparison of the Rabha fricatives are also made with the Korean fricatives to confirm the existence of the aspirated fricatives.

From the above discussions, it is seen that there is still a need to explore a novel framework that will effectively utilize the advantages of both linguistic and signal processing knowledge-based approaches for acoustic-phonetic analysis and automatic classification of aspirated fricatives and aspirated nasals. We have some additional specific goals in mind to address all the issues addressed above. Firstly, we will examine the acoustic properties of fricatives and the properties of nasals. In particular, we will investigate how and to what degree the two-way contrast of non-stops can be differentiated by the acoustic-phonetic features associated with aspiration in the Rabha fricatives and Angami nasals. Secondly, we will examine the aerodynamic properties of non-stops in terms of airflow and air pressure, which has been covered in this Chapter. There are no such studies for Rabha fricatives and only a few studies available that have examined the aerodynamic properties of nasals, such as airflow and air pressure of aspirated nasals. As far as we know, the current study is the first to analyze the acoustic-phonetic properties of Rabha fricatives. We also conducted a perception experiment to see whether the Rabha fricatives have two-way contrast based on aspiration, which is covered in Chapter 3. Thirdly, we will investigate the source features and extract relevant information to categorize the aspirated non-stop consonants from their unaspirated counterparts in Chapter 4. Fourthly, this study provides a cross-linguistic comparison across two very non-related

languages which is described in Chapter 5. By examining the acoustic properties of Korean fricatives using the same methods as for Rabha fricatives, we will determine what acoustic parameters differentiate /s/ and /s^h/, and how the Rabha /s/ is best classified. The present study is the first one directly comparing the acoustic properties of Korean fricatives with those of Rabha fricatives. Based on this comparison, we will provide phonetic evidence that the Rabha /s/ contrast is two-way based on the presence or absence of the aspiration category. While the previous chapters concentrate on the classification of aspirated-unaspirated fricatives and nasals as two separate tasks, the motive of Chapter 6 is to extract an acoustic parameter that would discriminate the aspirated fricatives as well as aspirated nasals from their unaspirated counterparts. Finally, Chapter 7 attempts to automate the process by utilizing all the acoustic-phonetic information to extract features using a multi-modal approach. The extracted features are further tested in an automatic speech recognition system.

3

Database, production and perception of aspirated fricatives and aspirated nasals

Contents

3.1 Database collection

In order to study the aspirated and unaspirated fricatives and nasals, the database was created for the Rabha and Angami as there is no standard database available in the public domain for those languages. The database preparation of Rabha is a part of this thesis work, while the Angami database has been collected from a research scholar of IIT Guwahati. Aspirated fricatives are also reported in the Korean language. Hence, we also collected the Korean speech database with Korean fricatives contrasting in aspiration as part of this dissertation.

In the current study, data was collected with the aim of capturing the two-way contrast of the fricatives and nasals of the languages under research. Differences in the manner of articulation will form the basis of the two-class classification of the non-stop consonants in this study. In order to avoid phonetic variations arising due to speaker and gender differences, the data will be normalized before conducting analyses on them. Before we proceed, a brief introduction to the languages studied in the thesis is provided in the following sections.

3.2 Languages in this study: Rabha, Angami, and Korean

India is a culturally and linguistically rich country with a population of 121 crore. A land of multiple language speakers, India is home to 121 languages each spoken by 10,000 or more people in India [109]. According to the Census of India 2011, there are a total of 121 languages and 270 mother tongues, of which 22 languages are specified in the Eighth Schedule of the Indian constitution. North-East India, which comprises eight states of India, consists only 7.9 percent of the country's total geographical area but is home to more than 75 % of the languages belonging to the four language families, viz Indo-Aryan, Tibeto-Burman, Austro-Asiatic, and Dravidian. Also known as the seven sisters, North-East India, is the most linguistically diverse region in the world. North-East India is home to about 200 Tibeto-Burman languages, which are under-documented and under-resourced. The phonology of these languages is very different from the Indo-Aryan languages spoken in the rest of India. For example, while the Indo-Aryan languages show rich phonation contrasts, often contrasting stops among voiceless, voiceless aspirated, voiced, voiced aspirated (breathy), the Tibeto-Burman languages restrict such phonation contrasts only to voiced

and voiceless aspirated stops [2]. Similarly, in the Tibeto-Burman languages, aspiration can be a contrasting feature for both affricates and fricatives.

Hence, in this thesis, we have tried analyzing the contrast between fricatives and nasals in terms of their aspiration. Rabha is chosen to study the aspiration contrast in fricatives, while to study the aspiration contrast in nasals, Angami language is chosen. In order to confirm our methods for analyzing and detecting aspiration in fricatives and nasals, we also considered using speech data from Korean language, which is well-known for its contrasts in fricatives in terms of aspiration. The following subsection gives a brief description of the Rabha language, followed by a description of the Angami language. Finally, the third subsection provides an overview of the Korean language.

3.2.1 Rabha

The Rabhas are one of the indigenous tribes of Assam, Meghalaya, and West Bengal. Linguistically, the Rabha language belongs to the Bodo-Garo sub-group under the Assam-Burmese group of the Tibeto-Burman languages [110]. According to Chatterjee (1974) [111–114], Rabha has many similarities to the Garo in particular. Joseph(2005) [114] states that the language also has many similarities to the Bodo and the Tiwa languages apart from the Garo language. But from the point of ethnic affinities, [2, 115, 116] they resemble the characteristics of the Tibeto-Burman linguistic groups. The Rabhas primarily inhabit the Goalpara and Kamrup districts of Assam and East and West Garo Hills district of Meghalaya, with a few population in other districts of Assam ,West Bengal, Nepal and Bangladesh. According to the census report of 2011 [109],the population of the Rabhas in India has been recorded as 1,39,986 of which a majority of 1,01,752 people reside in the state of Assam. There are 21,671 Rabha speakers residing in the state of Meghalaya. It is to be noted that according to the Census of India, there is a 15.04% decline in the Rabha speaking population from 2001 to 2011.

The Rabhas have some clan groups, which consist of Rongdani, Maitori, Koch (or Pani Koch), Hana, Pati, Dahuri, Total, Bitalia, etc. [117]. The first three groups are the major socio-linguistic groups. They maintain their inherent language and culture compared to the other minority groups. According to Rabha, 2010 [110], the language comprises of three dialects, i.e., Rongdani, Maitori, and Koch, showing some variations in the lexical and phonological level which make one group

separate from the other [110, 116].

Rabha is a tonal language as found in a number of Tibeto-Burman languages [118, 119]. The script of Rabha has been adopted from the Assamese script. Standard Rabha, closest to the Rongdani variety, consists of 8 vowels and 22 consonants, excluding one glottal stop and two semivowels(Rabha Khurangmuk 2015) in its phonemic inventory. In the case of consonant phonemes, Rabha has common similarities with the Boro language, another language of Assam of Tibeto-Burman origin. One such similarity is possibility of the presence of aspirated voiceless alveolar fricative as observed in the case of Bodo fricative [16]. The voiceless alveolar fricative, /s/, in Rabha can occur in initial, medial, and final positions of a word and can combine with all the vowels in the phoneme inventory [120]. An acoustic study on the Rabha fricatives revealed that the voiceless alveolar fricative has aspiration in its phonetic realization. The preliminary investigation will focus on the detection of aspiration in Rabha alveolar fricatives (s and s^h) in two varieties of Rabha, namely, Rongdani and Maitori.

3.2.2 Angami

Angami language, also known as Tenyidie, belongs to the Tibeto-Burman language family and is spoken mainly in the Nagaland state of the north-eastern part of India and some parts of Maharashtra and Manipur states. According to the census report of 2011, there are 1,52,796 Angami speakers in India, of which a majority of 1,51,883 speakers reside in the state of Nagaland [109]. With a population of 1,32,225 people who returned their mother-tongue as Angami in 2001, positive growth has been observed over the past decade.

According to the data in Ethnologue, Angami has many dialects: Dzuna, Kehena, Khonoma, Chakroma (Western Angami), Mima, Nali, Mozome, Tengima (Kohima). Among these, Tengima (Kohima) dialect is considered the standard dialect. Naga Chokri and Naga Khezha are eastern Angami groups with their own dialects. This study describes the phonetic inventory of one of the smaller Angami dialects, Khonoma, which no more than 5000 people speak in the extreme west of the Angami region. Khonoma Angami is also a tonal language with 4 level tones. The phonemic inventory of Khonoma Angami has 6 vowels and 45 consonants with a set of aspirated nasals and is written using Roman script [15, 22].

3.2.3 Korean

Korean is the national language of South Korea and North Korea and is spoken by 81,520,400 speakers worldwide [121]. Of this, 50,200,000 speakers reside in South Korea as per the census in 2020 [11]. There are six dialects spoken in South Korea, namely, Gyeonggi, Gangwon, Chungcheong, Gyeongsang, Jeollado dialect, and Jeju dialect. The Gyeonggi dialect is used as the standard Korean language among these dialects.

Hangeol, a unique alphabet, is used to write the Korean language. Korean has 19 consonants, 10 vowels. The consonant inventory of Korean can be represented in terms of their place of articulation and manner of articulation. As per the place of articulation, the Korean phoneme inventory has a three-way distinction between labial, coronal, and velar stops and a single position for affricates [122]. The Korean stops and affricates are represented by three-way contrast in manner of articulation, while the fricatives have two-way contrast contrasting in aspiration [35].

3.3 Participants

A total of 44 Rabha speakers, 19 Korean speakers, and 11 Angami speakers were recorded for the study. The Rabha participants consisted of the Rongdani and Maitori dialect speakers of the language. The Rabha speech dataset was recorded and collected from 24 villages of Lakhipur district, Assam and Tikrikilla district, Meghalaya. For Korean, the database was prepared in two phases. In phase I, 7 Korean speakers (5 male, 2 female) in the age group of 21-25 years, with an average age of 23 years, were recorded at the Phonetics and Phonology Lab, IIT Guwahati, in a sound-attenuated recording booth.

In phase II, Korean data was collected in Seoul, South Korea, from 12 native Korean speakers residing in Seoul for more than a decade at the time of data collection. While 9 were born and raised in Seoul, 3 reported that they were not originally from the Seoul area. Despite not being from Seoul, these three speakers spoke standard Korean, i.e., the Seoul dialect, confirmed by a native Seoul speaker. There were 11 female and 1 male speaker, in the age group of 25-39 years, with an average age of 31.7 years volunteered to participate in the study in phase II.

Speech data of 11 Angami speakers were recorded reading a list of words, and from that, the tokens having nasals were selected for analysis. For the acoustic study, three nasal consonants

Figure 3.1: Database collection procedure demonstrated for some of the native Rabha speakers

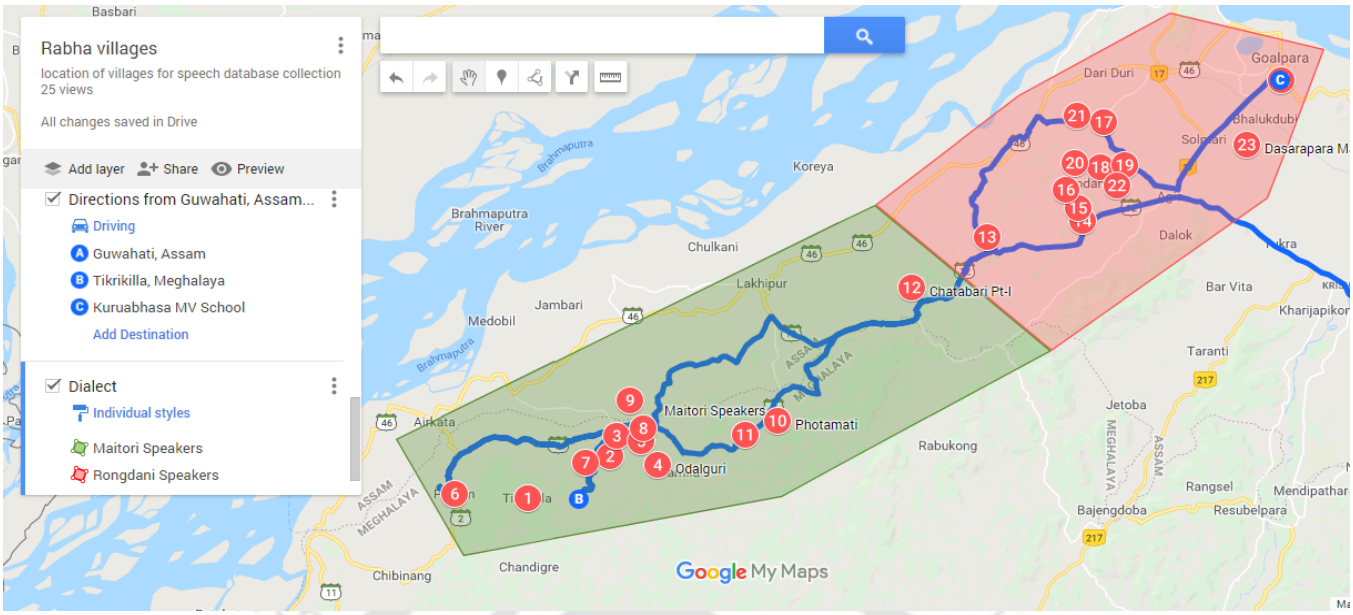

Figure 3.2: The Rabha database was collected from 24 Rabha villages of Assam and Meghalaya, India

Table 3.1: Summary of the subjects in the speech database preparation

Database	Rabha	Korean	Angami
Male	29	6	7
Female	15	13	4
Total	44	19	11

were examined. The Angami database had 121 tokens with 41 aspirated nasals and 80 unaspirated nasals. The 41 aspirated nasals have 6 minimal pairs with contrast in aspiration (See Table 3.4).

3.4 Speech Materials

The speech tokens for developing the Rabha, Korean and Angami database is described in detail in Table 3.2, 3.3 and 3.4. During the Rabha speech utterance analysis, it is also observed that aspiration tends to occur more when the fricative is followed by the low vowels in Rabha. The amount of aspiration increases as the following vowel decreases in height. Similar tendencies are also observed in Korean [34, 62, 63]. More aspirated fricatives are observed when followed by the low vowel /a/, whereas most unaspirated fricatives are followed by higher vowels such as /e/, /u/, and /i/. The emergence of the aspirated fricatives in Rabha can be traced back to its Proto-Tibeto-Burman (PTB) roots [1]. The onsets */ts/ and */dz/ in PTB have developed into /s^h/ in several Tibeto-Burman languages, such as, Burmese. In case of Rabha, aspirated [s^h] are more noticeable

Table 3.2: Target words for Rabha fricatives

Word	Gloss
[sʰa]	ache
[sa]	eat
[sak]	leaf
[sʰam]	grass, mortar
[sam]	wait
[sʰo]	burn, rot
[so]	mosquito
[seŋ]	waist
[si]	blood
[siŋ]	ask
[soŋ]	village
[su]	wound

in words with */ts/ onsets in their PTB cognates.

The Korean participants read out a list of words mentioned in Table 3.3. The text corpus included four minimal pairs contrasting in aspiration. The contrast in the two fricatives varied across vowel contexts, and the aspiration influence was found to be higher for low vowels than for high vowels [34, 62, 63]. Hence, our study chooses a Korean word list containing a post-fricative vowel /a/.

The Angami speakers read out a list of words which consisted of voiced and voiceless nasals in three places of articulation, namely, bilabial, alveolar, and palatal (m-mʰ, n- nʰ ŋ-ŋʰ). The target nasal sounds were monosyllabic (CV) and followed by five vowels (i, e, a, o, u).

The recorded tokens were manually categorized into aspirated and unaspirated classes which was necessary for training and modeling purposes. Based on auditory and spectrographic evidence, expert phoneticians manually annotated the speech utterances as aspirated/ unaspirated.

3.5 Recording

The target words in the speech corpus were recorded in three different environments and were repeated in isolation, in a fixed phrase, and in a sentence frame. Each set was repeated five times. The direction was given individually for each time at the beginning of a recording, and it took roughly 20 minutes per speaker to finish the whole task. As compensation, 500 Indian Rupees

Table 3.3: Target words for Korean fricatives

Word	Gloss
[s^{h}aɬ]	flesh
[s^{h}an]	mountain
[s^{h}am]	three or ginseng
[s^{h}aŋ]	prize or table
[saɬ]	rice
[san]	cheap or wrapped
[sam]	wrap(noun)
[saŋ]	couple, pair(numeral classifier)

Table 3.4: Target words for Angami nasals

Word	Gloss	Word	Gloss
[m^{h}a]	something to work	[n^{h}a]	plants
[ma]	people/price	[na]	blackseed/budding
[m^{h}i]	eye	[ŋ^{h}a]	messy
[mi]	light/fire	[ŋa]	ripe/crazy/tired
[m^{h}o]	above	[n^{h}o]	plastic cover
[mo]	cow mooing	[no]	child/suffocated/you
[ne]	push	[n^{h}u]	something to carry
[ni]/[nii]	earring/happy	[nu]	breastfeeding

and 5,000 Korean Won were paid to each Korean speaker participating in Phase I and Phase II, respectively. Each of the Rabha speakers who contributed to the preparation of the Rabha speech corpus were compensated as well.

All the utterances were recorded in a quiet environment with a Shure SM10 headset microphone attached to a Tascam DR 100 MKII solid-state audio recorder in .wav format. The speech samples were recorded at a sampling frequency of 44.1 kHz and manually annotated with PRAAT [4].

3.5.1 Speech database preparation

The speech samples were recorded with a sampling rate of 44.1 kHz and annotated manually with the PRAAT toolkit. For the analysis, the speech samples were resampled to 16 kHz. The samples are marked for the boundaries of nasals/ fricatives, aspiration, and vowel using auditory information and visual impression from the spectrographic display as shown in Figure 3.3 and 3.4. The collected database is from an under-resourced language, which is sufficient to give a glimpse of the overall phenomena for the knowledge-based classification analysis.

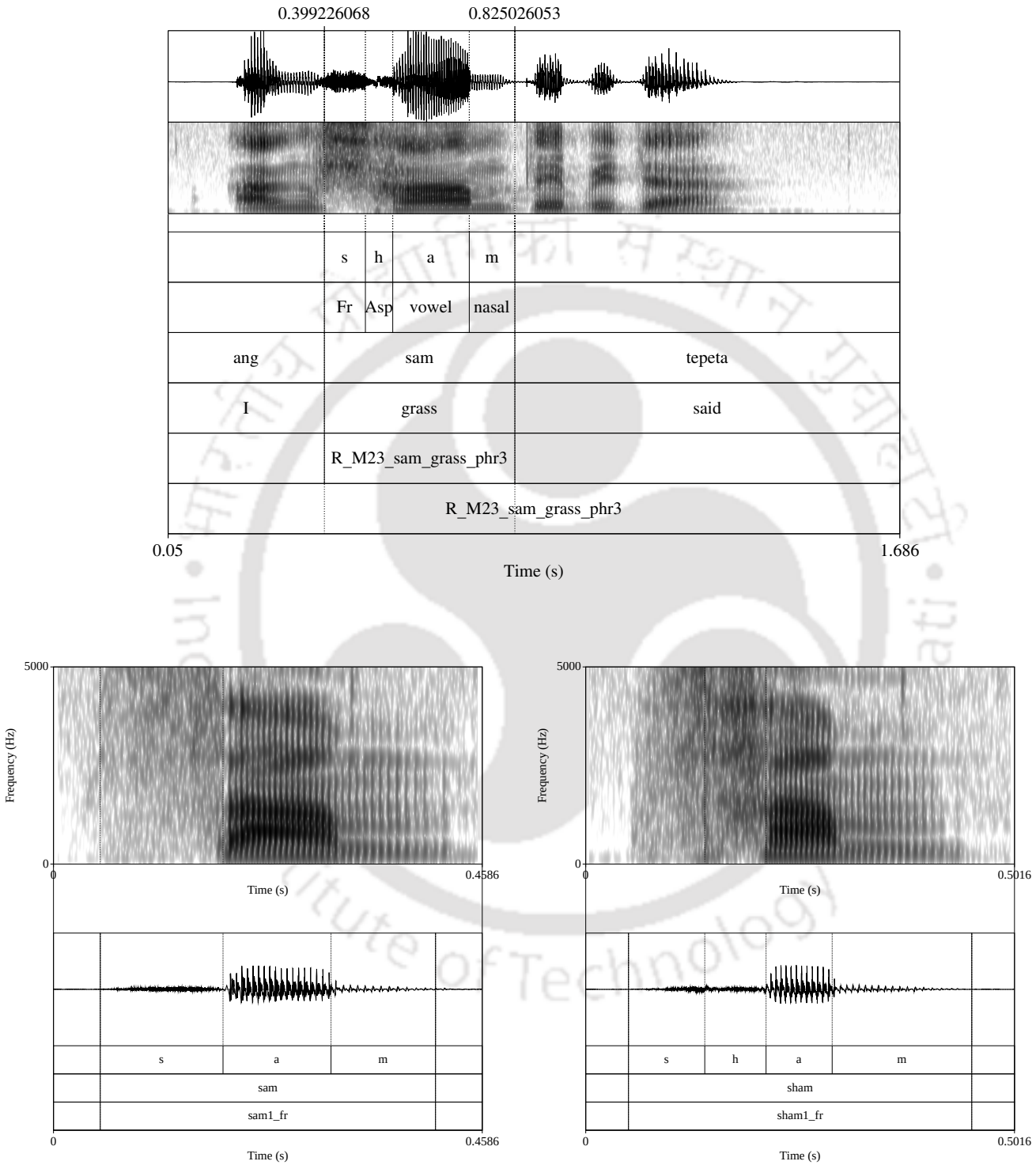

Figure 3.3: Spectrogram and manual annotation of Rabha fricative /sam/; /sam/ and /s^{h}am/ in Korean fricatives

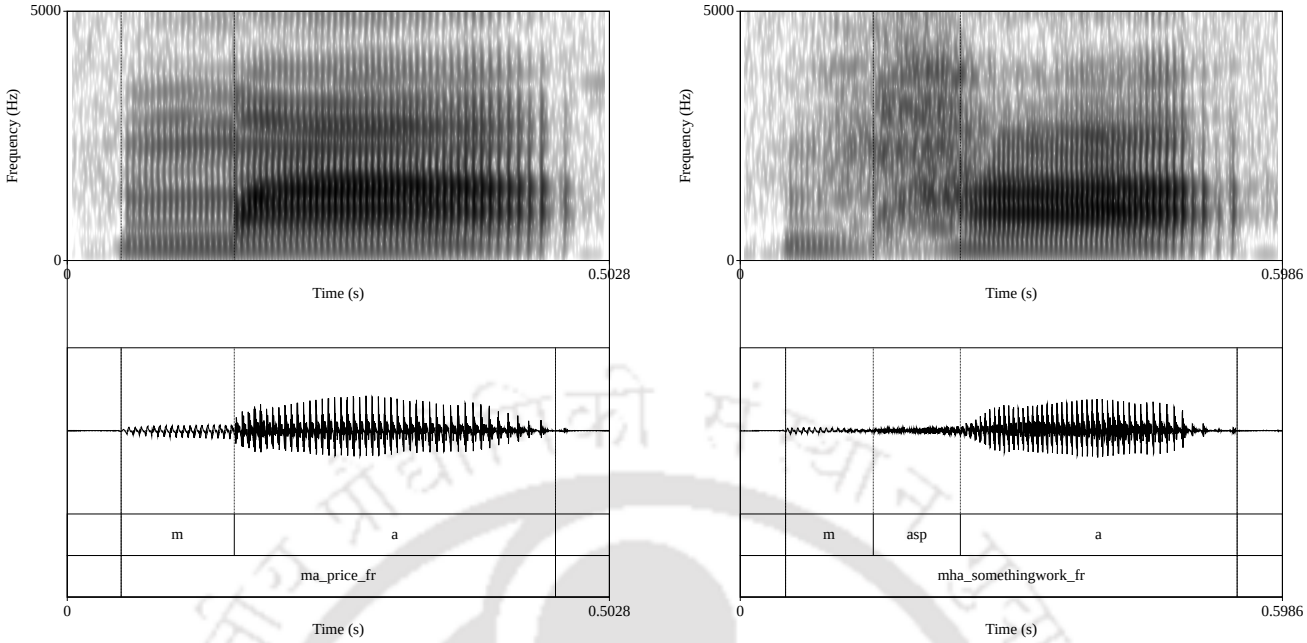

Figure 3.4: Spectrogram and manual annotation of /ma/ and /m^{h}a/ in Angami nasals

3.6 Perceptual evidence for aspiration contrast in Rabha fricatives

Korean fricatives are often classified as two-way contrast fricatives: aspirated and unaspirated. Even some previous study has hinted at the two-way contrast of the Angami nasals. However, no study has been attempted for the phonological fricative contrast in Rabha. As such, the phonemic and phonological status of the Rabha fricatives is not clear. Without any perceptual experiment in Rabha that verifies that aspiration has any sort of linguistic consequence, the production analysis and, therefore, the automatic classification of the Rabha fricatives would be meaningless. Hence, this perception experiment would give an insight into the existence or non-existence of phonemic pairs of Rabha fricatives, where aspiration may serve as a cue. A perception experiment was designed to test whether Rabha listeners distinguish the fricatives as aspirated and unaspirated in the same way as native Korean listeners. Two experiments were conducted in order to examine the degree of similarity between the minimal pairs of the fricatives. Experiment 1 was a Same-Different (AX) discrimination task that investigated listeners' perceived similarity judgments [28]. In Experiment 1, native Rabha speakers were asked to decide whether paired auditory stimuli containing the fricatives were the same or different. In a discrimination task, 58 stimuli pairs (37 unique stimuli) from three

Table 3.5: The Rabha speech stimuli considered for the identification test and their roots of origin [1, 2]. MoA: Manner of articulation, PTB: Proto-Tibeto-Burman, PST: Proto-Sino-Tibetan

word	gloss	tone	MoA	PTB
sa	ache	L	asp	-
	eat	H	unasp	-
sam	grass	H	asp	tswa/r-tsa-n(PST)
	mortar	H	asp	tsum
	wait	L	unasp	dzoŋ
so	burn	H	asp	m(t)-sik, tsik
	rot	H	asp	tswəy, zya:w, zyu(w)
	mosquito	L	unasp	kran

minimal pairs (/sa/,/sam/,/so/) of Rabha fricative and two minimal pairs (/sam/,/sal/) of Korean fricative were presented to 4 female and 12 male subjects. In an identification task, 28 stimuli from three minimal pairs (/sa/,/sam/,/so/) were presented to 3 female and 12 male subjects.

3.6.1 Subjects

15 native Rabha speakers participated in the perception experiment. There were 12 male 3 female who belonged to the 8 villages: Rampur, Tilapara, Borjuli, Darani, Bamundanga, Dairong of Goalpara district of Assam, India. The details of the Rabha speech tokens used in the perception experiment have been mentioned in Table 3.5.

3.6.2 Preparation of stimuli

It is important to include stimuli in which the sounds are identical in a perception experiment hence the stimuli were manipulated so that both sounds have identical

- Duration
- Tone information
- Intensity

All the above three parameters were kept identical for all the minimal pairs such that the only perceived difference is the presence or absence of aspiration in the production of the fricatives.

The perception experiment is designed with appropriate controls and stimulus design that will more narrowly implicate aspiration as the contributing factor in perceiving the difference between

Figure 3.5: One of the subjects participating in the perception experiment

these words. Manipulations were performed in PRAAT by resynthesizing using the overlap-add method.

After selecting the stimuli pairs, other sounds are removed from the stimuli such that only target utterance is extracted. The duration is normalized using the overlap-add method in PRAAT from the two minimal pairs, where the factor is calculated using the following equation.

$$Factor = (S_l - S_s)/S_s \tag{3.1}$$

where S_l, S_s are the longest and shortest duration (in ms) of the speech stimuli from the minimal pairs.

For the manipulation of pitch, one of the stimuli from the minimal pair is chosen, and the 1st and last pitch points of the stimulus in PRAAT are adjusted with the mean frequency value with a deviation of 15 Hz. The pitch manipulation process is repeated for the second speech stimulus of the minimal pair. Thus F0 of the following vowel of the paired stimuli have been made similar.

The intensity of the speech stimuli is modified identically using the scale intensity parameter of PRAAT.

3.6.3 Discrimination test

While aspiration contrast in Korean fricatives is well-attested [44, 52, 123], such is not the case for Rabha. In order to confirm the phonemic status of aspirated fricatives in Rabha, we conducted a pilot perception study on sixteen subjects, native speakers of Rabha, twelve male, and four female. The perception tests consisted of the /sa/-/s^{h}a/ pair where /sa/ is associated with "to eat" and /s^{h}a/ is associated with "to ache". The stimuli produced by four native Rabha speakers (2 male, 2 female) were used in the discrimination experiment. The subjects were asked to participate in a discrimination test where they had to choose whether the two stimuli were the same or different. The test also consisted of control pairs where both the stimuli were the same. Along with the Rabha speech corpus, Korean minimal pairs contrasting in aspiration were also used in the discrimination experiment. The purpose of adding the Korean stimuli was to test whether the Rabha speakers can perceptually distinguish the variation in aspiration in Korean fricatives.

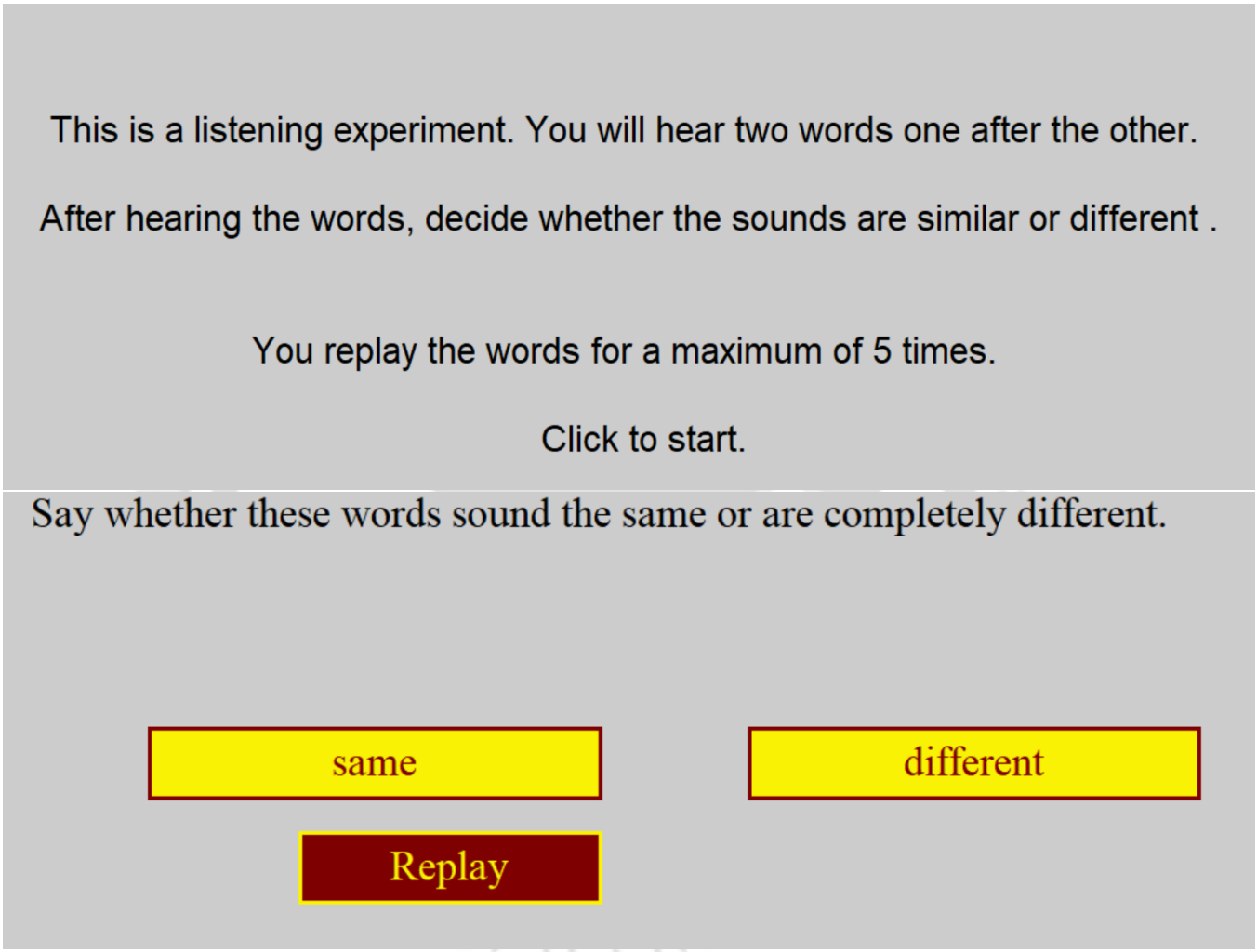

Figure 3.6: Interface for conducting the discrimination test of perception experiment

3.6.3.1 Methodology

Discrimination experiments are experiments in which a stimulus consists of multiple sub-stimuli played in sequence, and the subject has to judge the similarity between these sub-stimuli. The simplest discrimination task has only two sub-stimuli, and the subject has to say whether these are the same or different. It is important to include stimuli in which the sounds are identical in a discrimination experiment hence the stimuli were manipulated as mentioned in the previous section.

A 0.5-seconds silence is played before every stimulus so that the listener will not hear the stimulus immediately after his/her mouse click. The discrimination test presented syllables in pairs with an 800 ms inter-stimulus interval. Pair order was pseudo-random, 5 replications per stimulus, and the subject was provided a break after every 50 stimuli. Participants had to say whether the stimuli within each pair were the same (either two 'sam', to two 'so' syllables) or different (e.g. 'sam'–'s^{h}am'), according to their impression to hear the same thing or not. The experimenter registered the response by pressing one of two keyboard keys. The results of the perception experiments are analyzed using the D-prime value. D-prime (d') or discriminative index is a statistic used in signal detection theory to measure sensitivity or discriminability unaffected by response biases. D- prime is measured as the standardized difference between the mean of the signal present (in our case, difference in the two stimuli present) and signal absent (no difference in the two stimuli presented) distribution. The possible outcomes of the AX (same-different) experiment are as follows: "Yes" represents the presence of a signal or the difference to be detected, "No" represents the absence of a difference. The responses to the stimuli are categorised as Correct rejection, False alarm (F), Hit (H), and Miss. The hit rate is the proportion of trials where the stimulus was present, and the subject responded that the stimulus was present. The false alarm rate is the proportion of trials where the stimulus was not present, and the subject responded that the stimulus was present.

- **Hit rate H**: The probability of a yes response given the target is present. It is also represented as true positive (TP) in confusion matrix. $H = P(yes|present)$
- **False alarm rate F**: The probability of a yes response given the target is absent. It is also represented as false negative (FN) in the confusion matrix. $F = P(yes|absent)$

The d' value is calculated using the Macmillan and Creelman (2004)'s method as given in the

Table 3.6: An AX experiment as presented by Keating (2005).

	Response: Different (yes)	**Response: Same (no)**
Stimuli: YES (different)	HIT (TP)	MISS (FP)
Stimuli: NO (same)	FALSE ALARM (FN)	CORRECT REJECTION (TN)

equation below [124]:

$$d' = z(H) - z(F) \tag{3.2}$$

where $z(H)$ and $z(F)$ are the z transforms of hit rate and false alarm, respectively. The z-transform is calculated using $NORMSINV$ in Excel based on the standard normal distribution. It is to be mentioned that neither H nor F can be 0 or 1 (if so, the values are adjusted as demonstrated by Pat Keating [124]) Higher d' scores means higher sensitivity in discriminating the contrasts.

3.6.3.2 Results

The responses of the speakers were measured using the d' score [124]. As observed in Table 3.7, the participants could discriminate between the /sa/ and /s^{h}a/ with high D-prime values for the Korean data, confirming their ability to perceive aspiration contrast. Results of the discrimination test for the /sa/ and /s^{h}a/ pair of Rabha data show that the participants were sensitive to the distinction between the two sounds. The participants, when grouped based on their age difference into young and old, young group correctly discriminated between the /so/ and /s^{h}o/ pair of Rabha data with 73% correctness accuracy. For the /sa/ and /s^{h}a/ pair, the old group correctly discriminated 86.8 % while the young age group correctly discriminated 81.4% times. For the /sam/ and /s^{h}am/ pair, the old group correctly discriminated 73.9 % while the young age group correctly discriminated 77.2% times. Two participants had zero value, which indicates that participants could not distinguish between the two sounds and hit the "Same" button for all the cases.

3.6.4 Identification test

In identification tests, participants were presented with a sound and were asked to choose the correct meaning of the word they heard from a list of options. The experiment aimed to check if the participants were able to correctly identify the words presented to them.

Table 3.7: Results of the discrimination test obtained from the Rabha listeners. Here, d'_R, d'_K represents d' values for Rabha, Korean stimuli, respectively, while $d'_{R,K}$ represents d' values for combined stimuli of Rabha and Korean

Subject	Age	d'_R	d'_K	$d'_{R,K}$	average
F1	39	2.22359	2.765988	2.263262	2.494789
F2	26	2.478802	2.350416	2.468394	2.414609
F3	36	1.492381	1.172566	1.423576	1.332473
F5	30	2.500576	1.593423	2.365203	2.046999
M1	39	0.66629	0	0.466303	0.333145
M2	45	0.225636	0	0.208997	0.112818
M4	55	2.028775	0.952267	1.77669	1.490521
M5	48	2.129806	1.172566	1.952494	1.651186
M6	57	2.218019	1.593423	2.047367	1.905721
M7	73	1.16348	1.382994	1.20448	1.273237
M8	61	2.124235	2.765988	2.123344	2.445112
M9	43	3.177768	2.765988	3.116078	2.971878
M10	54	1.382994	0.220299	1.182652	0.801647
M11	49	1.384569	1.641911	1.434081	1.51324
M12	54	0.952267	1.382994	1.034238	1.16763
M13	42	1.438542	1.593423	1.463235	1.515982

3.6.4.1 Methodology

The preparation of the stimuli included speech utterances from 4 native Rabha speakers (2 Male, 2 Female). The stimuli were presented to the listeners by repeating each utterance 5 times in a random order (for 1 listening participant, the repeated utterances were reduced to 4 times due to her inconvenience).

A 0.5-seconds silence is played before every stimulus so that the listener will not hear the stimulus immediately after their mouse click. In the identification test, syllables were presented 5 replications per stimulus, and the subject was provided a break after every 40 stimuli. A total of 28 unique stimuli were presented to the participants for the listening test. Participants had to identify a single sound (e.g., Rabha /sam/) as one element of a set of categories, whichever closest they perceive among the set of meanings provided (grass, wait, mortar). Among the pictures presented in the identification task, the subject chooses the most similar picture to what they heard. The results of the perception experiments are analyzed using the value of correctness. The correctness is analyzed by comparing the meaning of the particular stimuli with that of the response. If the meaning of both stimuli and the response is the same, the result is recorded as correct, otherwise incorrect.

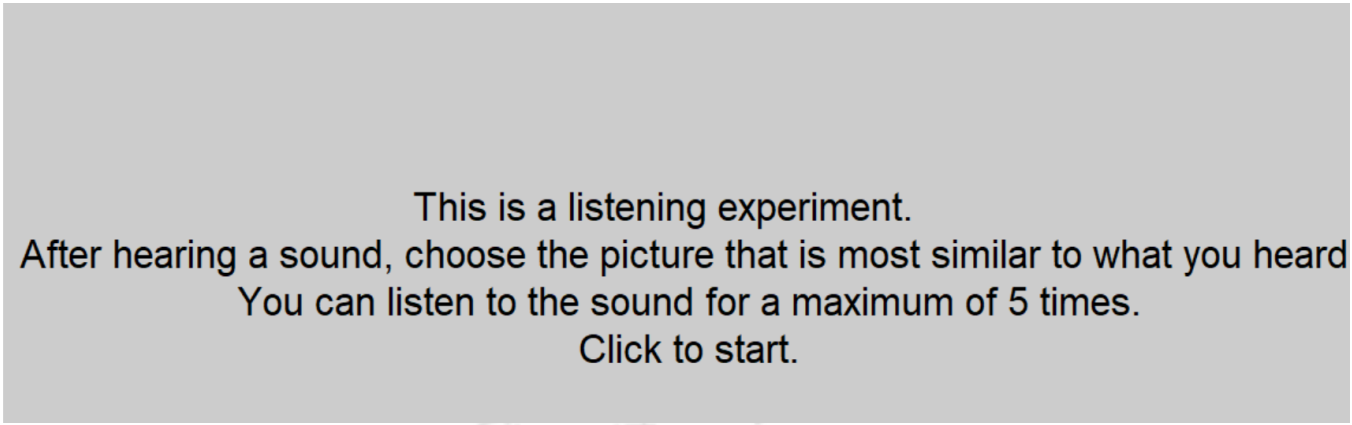

Figure 3.7: Interface for conducting the identification test of perception experiment

Table 3.8: Results of the identification test obtained from the Rabha listeners

Stimuli: sa			**Identification %**	
			F07	**M12**
	Age	**Subject**	**correctness**	
young	30	F5	100.00	75.00
	36	F3	70.00	80.00
	39	F1	100.00	100.00
	42	M13	50.00	70.00
	43	M9	100.00	80.00
old	48	M5	100.00	90.00
	49	M11	70.00	50.00
	54	M10	90.00	100.00
	54	M12	100.00	90.00
	55	M4	100.00	90.00
	57	M6	90.00	100.00
	61	M8	80.00	60.00
	73	M7	60.00	90.00

3.6.4.2 Results

Following the discrimination test, the subjects were asked to participate in an identification test where the stimuli were presented along with two images representing the meanings of the words. The participants had to identify the stimuli and choose one of the images provided.

Out of the 16 speakers that participated in the discrimination test, one female speaker refused to participate in the identification test.

Out of all the stimuli presented, 276 were considered for the perception results of /sa/ token. Also, the results of two male speakers have been omitted from the analysis as they did not fare

Figure 3.8: Choices for each stimuli of the identification experiment

Table 3.9: Percentage of correct identification of stimuli /sam/ and /so/ produced by 3 Rabha speakers and analysed by 14 Rabha listeners

Stimuli	**sam**	**Stimuli**	**so**
F07	66	F15	57
F15	55	F16	71
M23	67	M24	63
correctness %	63	**correctness %**	64

well in the discrimination test. For the word /sam/ and /so/ stimuli, 621 and 414 stimuli were considered for analyzing the results. Results of one male speaker were omitted from the analysis as the participant did not perceive the stimuli properly and wrongly identified most of the cases. As seen in the Table 3.8, when the /sa/ stimuli were presented to the participants, they correctly identified the word with their corresponding meaning with an average of 82.68%.

Regarding the identification test, the 14 Rabha listeners could correctly associate the stimuli with the meaning with an average accuracy of 63% and 64% for the utterance /sam/ and /so/, respectively (refer 3.9). The varied responses for these two speech utterances indicate that the participants were confused while correctly assigning the responses to the stimuli. The reason for the varied response may be because these two stimuli have 3 meanings, unlike the two different meanings for the word /sa/. Hence, comparatively, there is more chance of confusion in perception when the word is associated with three meanings. Besides the confusion, the listeners still perform above the chance level, and therefore, this confirms that the Rabha fricatives are phonemically distinguished in terms of aspiration.

3.6.5 Discussion

We are not aware of any studies that have investigated the perception of Rabha fricatives by listeners of any dialect. It is possible for Rabha fricatives to be neutralized in perception even if they are not neutralized in production. One such case of perceptual neutralization among Korean dialect speakers has been reported earlier [62]. Even if there are acoustic differences between the categories, it is possible for the categories to be perceptually neutralized. That is, listeners may still perceive the categories as the same even if there are acoustic differences between them.

Further systematic perception tests are required to know exactly which acoustic/perceptual

cues embedded in vowels are utilized by the Rabha listeners for the discrimination task of the fricatives contrasting in aspiration. For example, while aspirates may be associated with higher tones in one language, they may be associated with lower tones in another language. Following the perceptual analysis conducted in this section that reports phonemically distinguished Rabha fricatives, production analysis of the Rabha fricatives will be studied in the next section.

3.7 Acoustic Analysis of fricatives and nasals

3.7.1 Production Analysis

As mentioned earlier, a phonological distinction between unaspirated /s/ and aspirated /s^h/ has been observed in the Rabha language [43]. Also, the nasals in Angami have shown contrast based on aspiration, thus contributing to /m/- /m^h/, /n/-/n^h/ and /ŋ/-/$ŋ^h$/ pairs in the language [43]. However, unlike Korean, that has some literature available on the acoustic qualities for its aspirated fricatives [14, 44, 51, 52], aspirated fricatives in Rabha and aspirated nasals in Angami are rarely described [43, 44, 125]. Hence, in this work, we conduct an acoustic study on aspiration in the widely documented aspirated fricatives in Korean and compare the study with the aspiration in the rarely documented aspirated fricatives and aspirated nasals in Rabha and Angami respectively. The motivation is to check whether the behaviour of the aspiration features in the aspirated fricatives and aspirated nasals in Rabha and Angami, correlates with that of the Korean aspirated fricatives. In order to do that measures described in previous studies for Korean and Bodo aspirated fricatives are considered and their effectiveness in distinctly characterizing Angami and Rabha aspiration is explored [16, 44].

This paper will try to analyze the characteristics of the aspirated nasal by studying its features and comparing them with the characteristics of the aspirated fricatives. The goal is to check the similarity or dissimilarity of the acoustic features of aspiration in the rarely studied aspirated nasals with that of the aspiration in the previously studied aspirated fricatives [16, 44, 52]. Since not many studies have been done to understand the aspirated nasals, this effort will help in understanding the characteristics of the aspiration in the aspirated nasals. There is a two-way contrast in the nasals of Angami as well as in the case of Rabha fricatives - aspirated and unaspirated. The focus of this study is the phonetic analysis of the Angami phonemes /m/, /n/, /ŋ/ and /m^h/, /n^h/, /$ŋ^h$/

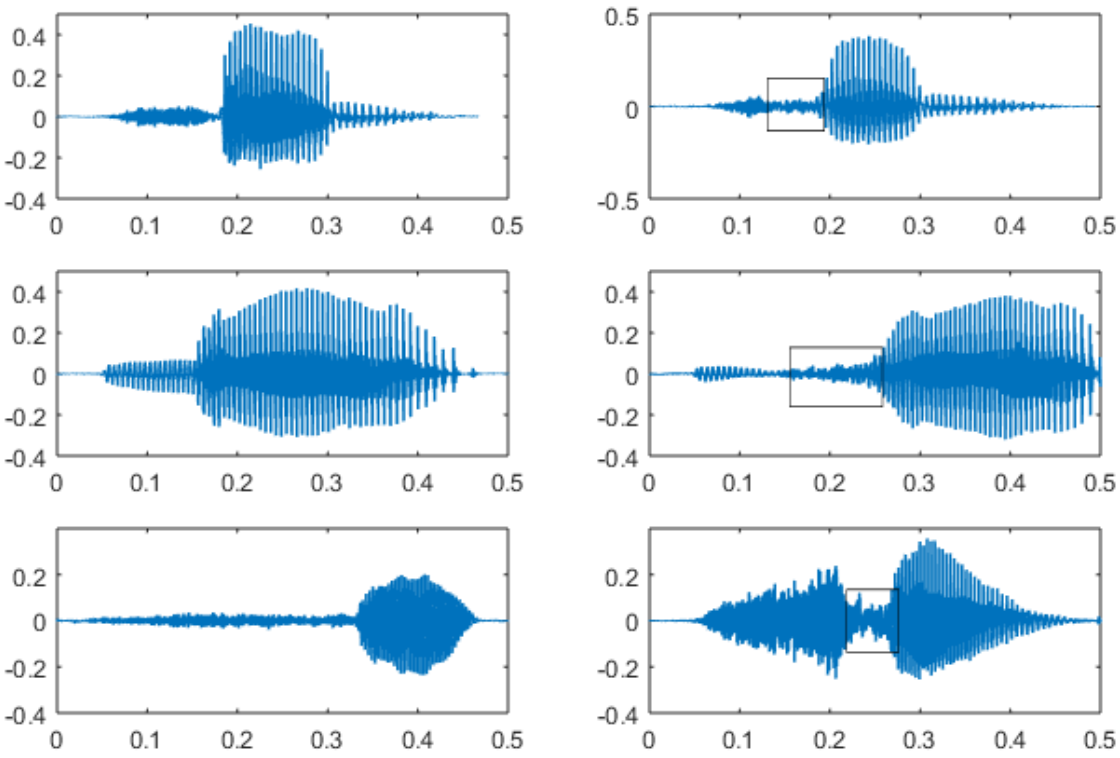

Figure 3.9: Waveforms of the speech signal (from top to bottom) (a) Korean: /sam/, /s^{h}am/(b) Angami: /ma/, /m^{h}a/ and (c) Rabha: /suk/, /s^{h}a/ (the black box highlights the tentative aspiration region)

having different place of articulation (i.e labial, alveolar and velar origin respectively). Although a detection mechanism for aspiration in Rabha alveolar fricatives has been proposed [36], several acoustic measures suggested in the previous literature were not used. Since the alveolar fricative in Korean too has two-way contrast - /s/ and /s^{h}/ i.e., unaspirated and aspirated counterparts, hence an acoustic comparison of the Rabha fricatives are also made with the Korean fricatives to confirm the existence of the aspirated fricatives. As there are limited studies on aspirated nasals, in this chapter, we try to investigate the properties of aspirated nasals by comparing it with the aspiration in the widely studied Korean aspirated fricatives. The following section discusses the methodology for the acoustic analysis of the consonants reported in the study. Further, it summarizes the acoustic analysis results of aspirated nasals and aspirated fricatives.

Previous literature [14, 32, 33, 44, 52, 126] has suggested several spectral and temporal measurements for distinguishing aspirated and unaspirated consonants. One of the measurements is duration [51]. The aspirated consonants tend to have a longer duration as compared to unaspirated

ones [44, 125]; however, when aspiration is excluded, it is much shorter than the unaspirated fricative [52]. According to Chang [44], a higher F1 onset is associated with longer VOT (Voice onset time) and lower F1 onset with shorter VOT. Also, the spectral tilt measure(H1-H2) is significantly higher for the aspirated fricative than for the unaspirated fricative. Since these measurements are seen to distinguish between aspirated and unaspirated fricatives in Korean and Bodo, a Tibeto-Burman language spoken in Assam, India [16, 52], hence an effort has been made to exploit these measurements for distinguishing between aspirated and unaspirated nasals in Angami as well. Apart from these three measures, i.e., duration, F1 onset, and spectral tilt, the center of gravity (COG) and intensity are also investigated to understand the characteristics of the Angami nasals and Rabha fricatives [89].

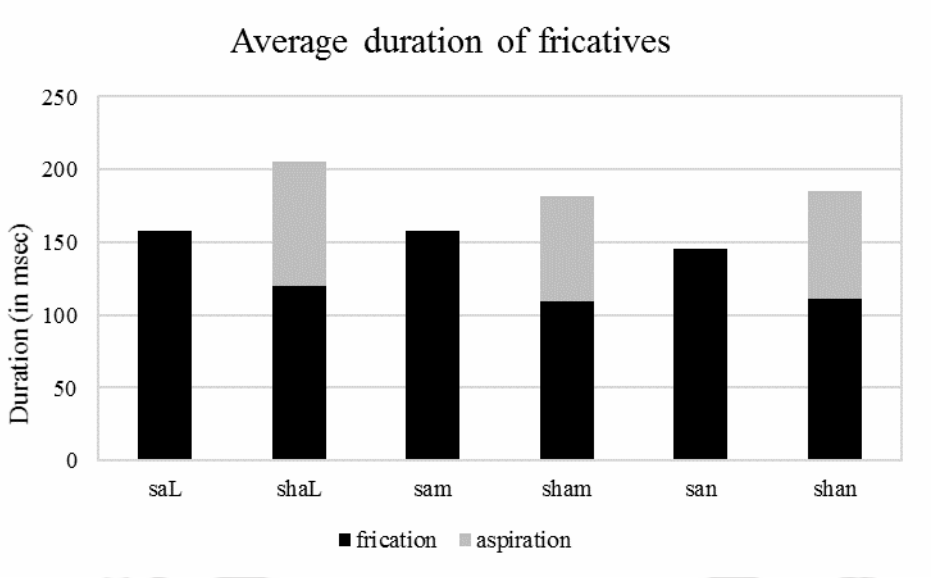

Figure 3.10: Average duration of Korean fricatives in the consonant segment

3.7.2 Acoustic cues

In this section, the results of all the measurements used to investigate the acoustic properties of nasals and fricatives are described. For the comparison, all words starting with /m/, /n/, /ŋ/ and /m^h/, /n^h/, /$ŋ^h$/ are averaged separately and analyzed further. Similarly,the fricatives /s/ and /s^h/ are categorised separately for further analysis. Since the Rabha fricatives were collected from two subgroups - Maitori and Rongdani hence, for the analysis, the data were categorized as Maitori unaspirated (MU), Maitori aspirated (MA), Rongdani unaspirated (RU), and Rongdani aspirated

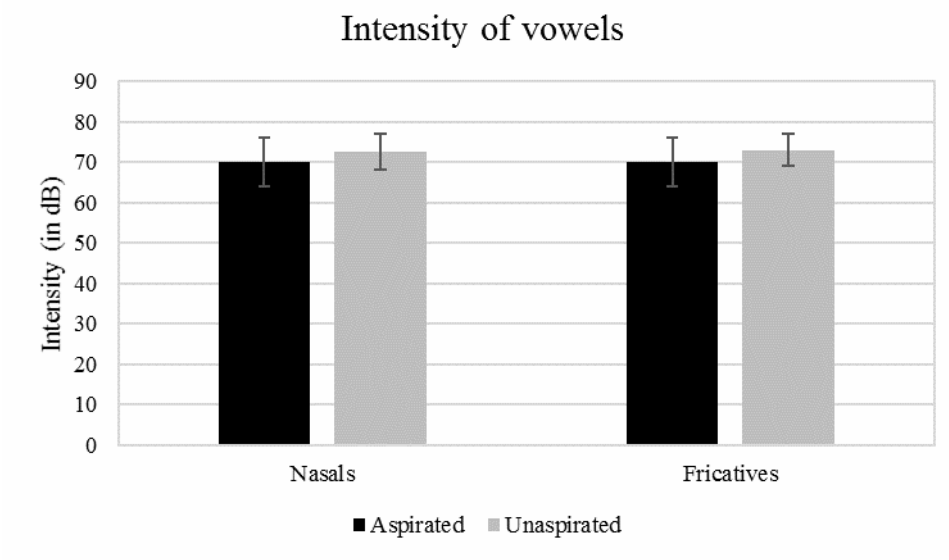

Figure 3.11: Intensity of vowels following the Angami nasals and Rabha fricatives

(RA). Finally, a combined study of all the measurements is done on all the three categories,i.e., Angami nasals, Korean fricatives, and Rabha fricatives which are further divided into aspirated and unaspirated categories.

3.7.2.1 Duration

Three durational measurements are made: vowel duration, nasal duration, aspiration duration in case of Angami nasals and frication, aspiration and vowel duration in Korean and Rabha fricatives. As seen in Fig 3.10, the aspirated consonants are longer than unaspirated ones, but in the case of vowels, the following vowels after the aspirated consonants are shorter in duration. In the case of the vowels following Korean and Rabha fricatives, the same pattern is observed. In the distribution plot, the trend for all the three categories, i.e., Angami nasals, Korean, and Rabha fricatives, is similar. Therefore only a single plot is shown in Fig 3.14 (f) for reference. As seen in Fig 3.14 (f), the overlap area of the aspirated and unaspirated curve is observed to be high. This suggests that duration is not a great acoustic measure to differentiate between the aspirated-unaspirated classes of fricatives and nasals used in this study.

3.7.2.2 Intensity

The intensity of the vowels following the nasals and fricatives is measured. It is seen that there is a very slight variation between the aspirated and unaspirated counterparts (see Fig 3.11). In the case of the nasals /m/- /m^{h}/ and /ŋ/-/ŋ^{h}/, the vowels after the unaspirated nasals have higher intensity than the unaspirated ones. However, for the nasals /n/ and /n^{h}/, it is observed that the trend in the intensity measurement is the opposite. In the case of Korean and Rabha fricatives, all the vowels following aspirated fricatives have a lower intensity value than their unaspirated counterparts. The intensity of the consonant region of the nasals is also examined. It is observed that the intensity in aspirated nasals too follows the same trend as in the case of vowels following the nasals, i.e., the unaspirated consonant region has a higher intensity than the aspirated ones. The distribution plots of the intensity values for the aspirated and unaspirated counterparts show greater overlap. Hence, the intensity is not considered a distinctive acoustic cue to distinguish between aspirated and unaspirated cases studied in this chapter.

3.7.2.3 Centre of gravity

The spectral center of gravity (COG) is a spectral parameter to measure the weighted mean of the frequencies in a spectrum. In other words, spectral COG indicates where the center of mass of the speech spectrum is located [127]. The COG values of aspirated consonants differ significantly from the unaspirated counterparts. For the fricatives (both Korean and Rabha), the COG is high for the unaspirated consonants (as seen in Fig 3.12), while for the Angami nasals, the COG value tends to be higher for the aspirated nasals. The presence of noise-like characteristics of aspiration in the nasals can be described as the reason for the opposite trend in the case of nasals. The distribution plot of COG values for the Angami nasals, Korean fricatives, and Rabha fricatives are shown in Fig 3.14 (a), (b), (c) respectively. Two distribution curves, i.e., one for the aspirated case and the other for the unaspirated case, are observed with lesser overlap for all the three studied language categories (see Fig 3.14 (a), (b), (c)). This suggests that COG as an acoustic parameter can differentiate between the aspirated-unaspirated counterparts.

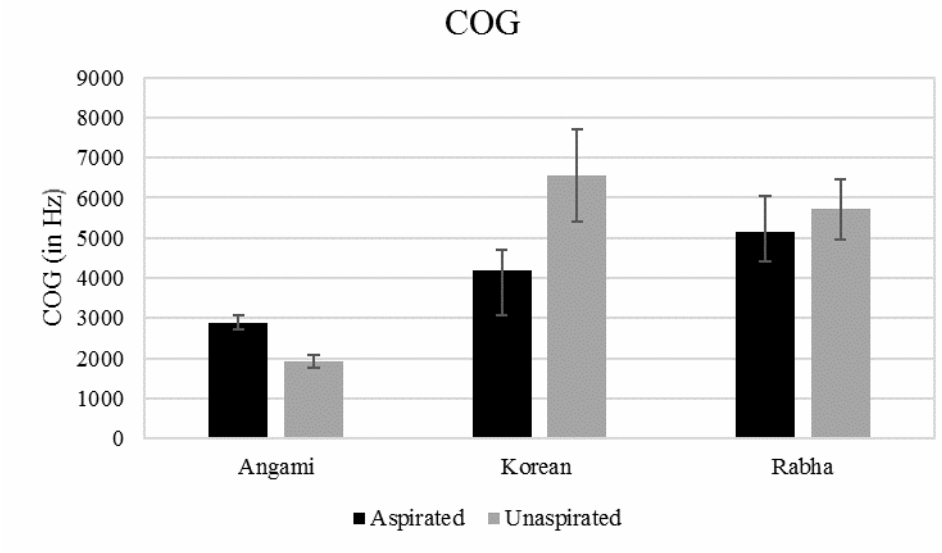

Figure 3.12: COG measures constructed from the Angami nasals, Korean fricatives and Rabha fricatives respectively

3.7.2.4 Spectral tilt

The H1-H2 value is a measure of spectral tilt, and it is found by taking the difference in amplitude between the first harmonic (H1) and second harmonic(H2). This measure has been associated with breathiness information. A greater H1-H2 value indicates the breathiness of the following vowel, whereas a smaller or negative value indicates pressed voicing quality [52]. A PRAAT script is run, which makes these measurements at different portions of the interval, based on the amount of chunking specified (three in our case). This parameter measures the f0 information at the locations of the first three formants. The spectral tilt is high for the vowels following aspirated nasals than the unaspirated counterparts indicating that the vowels after the aspirated consonants are breathy (or aspirated) in nature. The same trend is observed in the case of Korean and Rabha fricatives; in other words, aspirated fricatives have a higher spectral tilt than their unaspirated counterparts.

3.7.2.5 F1 onset

The F1 onset is the onset of the first formant (A) information and the measure is used as an acoustic correlate of aspiration [128]. It is measured at the first visible glottal cycle of the vowels following /m^{h}/, /n^{h}/, /ŋ^{h}/ and /s^{h}/. From Fig 3.13, it can be said that the F1 onset of the vowels following the unaspirated consonants starts lower than the aspirated ones for all the three studied

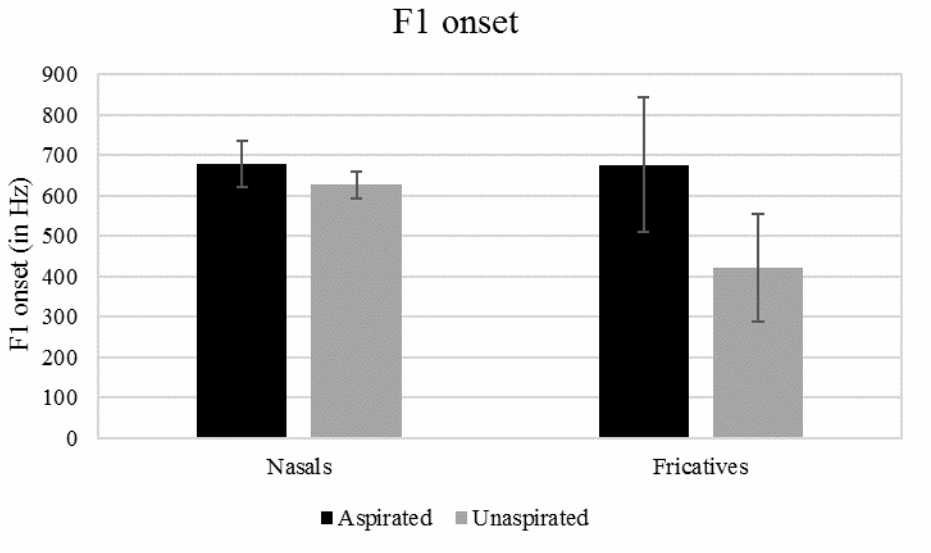

Figure 3.13: F1 onset of vowels following the Angami nasals and the Rabha fricatives respictively

cases, i.e., Angami nasals, Korean and Rabha fricatives. The pattern for F1 onset evidence is observed to be similar for Rabha and Korean fricatives; hence, only the Rabha case has been shown in the figure. As seen, the distribution plots corresponding to F1 onset values of the nasals and fricatives show greater overlap as compared to that between the distribution plots constructed from the COG evidence (refer to Fig 3.14 (d), (e)). Hence F1 onset parameter may not always be a distinctive cue to discriminate sounds in terms of aspiration present in the aspirated consonants.

3.7.3 Discussion

In both cases, i.e., aspirated fricatives and aspirated nasals, the aspiration has the formant characteristic of /h/, which is velar in pronunciation. The aspiration marked region perceptually sounds like /th/ or /h/ followed by the vowel sound. In Angami nasals, the formants corresponding to the aspiration feature occur during the entire hold of the nasal consonant.

This study confirms that Angami nasals have aspirated categorization and some of the trends such as spectral tilt, durational information are similar to Korean fricatives. Though the intensity does not give much information, the high F1 onset of the vowels following aspirated nasals supports aspirated analysis. The significantly lesser overlap in the distribution plot of COG evidence values of all the three studied categories suggests that the COG parameter can be used as a distinctive cue for distinguishing aspirated and unaspirated varieties. The aspirated groups of fricatives are

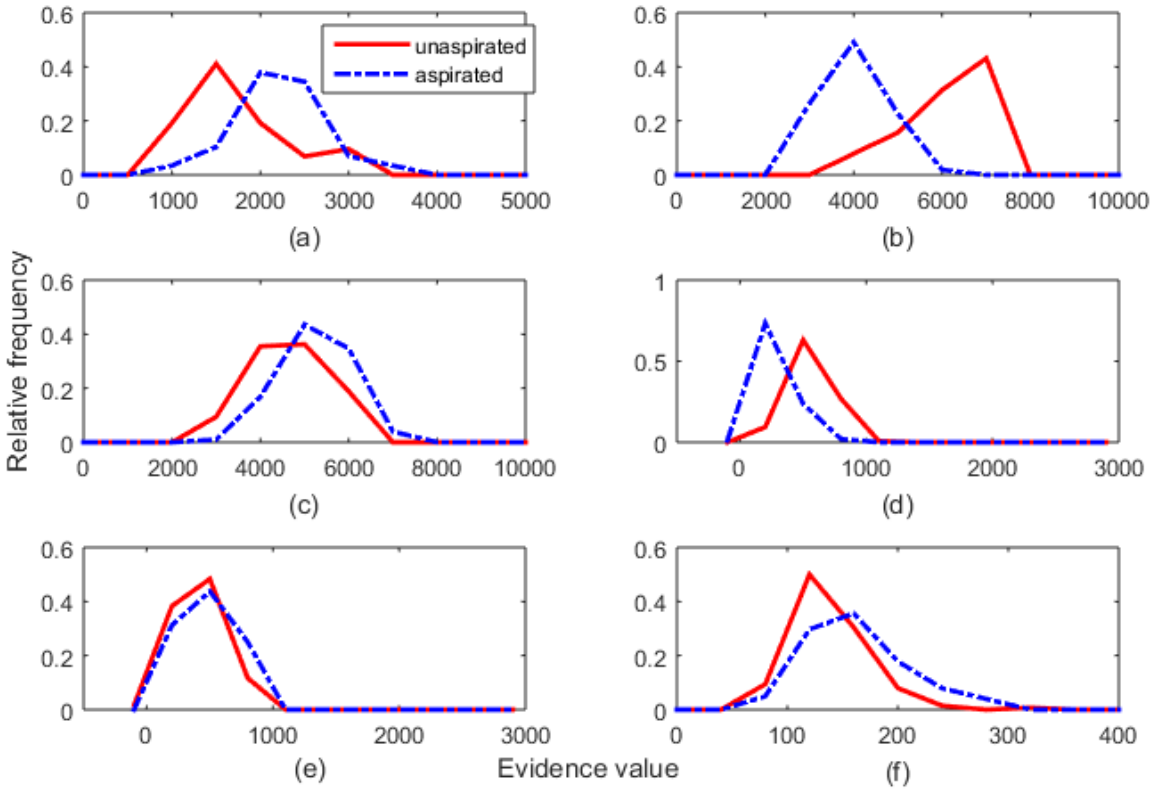

Figure 3.14: Distribution plot of COG constructed from (a) Angami nasals (b) Korean fricatives (c) Rabha fricatives; distribution of the F1 onset measure constructed from (d) Rabha fricatives and (e)Angami nasals, (f) distribution of the duration of the vowel following the Angami nasals

seen to have shorter vowel duration, lower centroid frequency, and higher F1 onset for the following vowel compared to the unaspirated groups of Rabha fricatives. Also, it is observed that the Rabha fricatives tend to have the highest amount of aspiration when followed by the low vowel /a/, whereas the least amount of aspiration is seen when followed by higher vowels such as /e/, /u/, and /i̥/. This study confirms that Angami nasals and Rabha fricatives have two contrast varieties in terms of aspiration.

The current chapters present the production analysis as well as the perceptual analysis of the aspirated fricatives. In the subsequent chapters, the signal processing method will be explored by utilizing the knowledge from the current branch for the automatic detection of aspirated fricatives and aspirated nasals.

4

Analysis of excitation source characteristics in aspirated fricatives and aspirated nasal consonants

Contents

4.1 Introduction

Aspiration is a distinctive feature in voiceless stop consonants in several languages, such as, Hindi, Urdu and Bangla [33]. The PHOIBLE database reported the existence of aspirated stop consonants, /p^h/, /t^h/ and /k^h/ in 363 of 2155 languages archived in the database [12]. On the other hand, occurrence of aspiration in non-stop consonants is quite rare. Only 6 out of the 2155 languages reported in the PHOIBLE database, confirmed the existence of aspirated non-stop consonants in their consonant inventories. The database reports the occurrence of phonologically aspirated /m/, /n/ and /s/ in 6 languages and the occurrence of aspirated /ŋ/, /f/ and /ʃ/ only in one language [12]. Reports of occurrence and studies concerned with the the acoustic characterization of non-stop aspirated sounds are scarce as they occur in some of the least studied languages of the world, such as, Angami, Bodo, Burmese, Sgaw Karen, Pumi, Qiádōng Hmongic, Central Mazahua, Cross River Mbembe, Barbareno and Embera-Catio [13–16]. On the other hand, Korean language is one of the well documented and well researched languages that also has aspiration in non-stop consonants. Languages, such as Korean or Pumi, have contrast between /s/ and /s^h/ [14]. In Bodo, the /s^h/ appears consistently, however, it does not contrast with /s/ in the language [16].

From an articulatory point of view, aspiration is considered a puff of air between the articulation of a consonant and the start of voicing [6]. It is characterized by wide-open glottis resulting in a breathy airflow. Aspiration is also characterized as a type of 'turbulant noise' resulting from a steady air flow from the glottis, exciting the resonances of the vocal tract. The excitation results in a whispery, breathy and voiceless portion of speech [3].

In case of aspirated stops, the period of voicelessness, following the release of a stop, serves as a distinct acoustic feature. This is accompanied by a distinctive burst, that provides sufficient acoustic evidence for identifying the boundary between the aspiration and the burst [58]. However, in case of the non-stop aspirated fricatives, acoustic features are not very distinctive. For example, in case of an aspirated nasal, lack of burst makes the detection of the boundary between the nasal portion and the aspiration part difficult. Similarly, distinguishing between an aspirated and an unaspirated nasal is also difficult due to the presence of strong nasal resonance in the higher frequency bands for both types of nasals [59]. Moreover, as the source for both frication and aspiration is turbulent noise, identifying the boundary between the aspirated portion and the frication portion of an aspirated

fricative also poses a challenge. In case of all types of fricatives, determining the boundary between the noise and voicing is challenging as there is no categorical boundary between the two [40]. The overlapping area between noise and voicing is also recognized as a distinct transient area in the literature [60]. As discussed earlier in previous chapter, while the boundaries of the aspiration region in the aspirated consonants can be determined relatively easily, the same is not the case for aspirated nasals and aspirated fricatives.

However, several distinguishing acoustic features contrasting an aspirated non-stop and its unaspirated counterpart are proposed in earlier works. In Sgaw Karen, aspiration in the aspirated fricatives is found to be significantly shorter than that in the aspirated stops [29]. Unlike in aspirated stops, aspirated non-stops have shorter voice onset time (VOT), which, also makes it harder to perceive it distinctly from its unaspirated counterpart [29]. It is also observed that, in aspirated fricatives, the opening of the glottal configuration is larger and the duration of aspiration is longer when compared to their unaspirated counterparts [44].

In case of Korean, the F0 onset of the vowel following an aspirated fricative is generally lower than the one following an aspirated stop. However, this is not statistically significant when compared to the F0 distribution of unaspirated and aspirated plosives [41]. Though high H1-H2 values are characteristics of Korean aspirated fricatives, as seen in [41], the same measure did not show any consistent pattern in the aspirated fricatives of Bodo [16]. The aspirated fricatives in Korean have longer VOT and are associated with higher F1 onsets and flatter F1 transition patterns, while the unaspirated fricatives are associated with shorter VOT, lower F1 onsets and steeper F1 transition patterns [44]. With longer VOT, the tongue has more time to move from consonant to vowel configuration, hence it results in a flatter F1 transition pattern. However, in case of unaspirated fricatives, due to short VOT, the tongue has little time to move from consonant to vowel position, resulting in steeper F1 transition pattern [44]. The results from the previous studies prompt us to look more closely into periodicity, vocal fold configuration and vocal tract constriction patterns, in order to detect aspiration in the non-stop consonants.

Aspiration results in contrastive phonemes in several languages. Hence, in a language that has a phonemic contrast between a pair of aspirated and unaspirated speech sounds, automatic identification of the aspiration feature is crucial in its phoneme recognition system. While phonemically

contrasting aspirated and unaspirated stops are fairly common in languages, only a handful of languages have phonemic contrasts between aspirated and unaspirated nasals and fricatives. Except Korean, most of the languages that have phonemic non-stop aspiration are low-resource languages that lack relevant speech data and detailed acoustic analysis. Considering this, in this work we explore the possibility of using spectro-temporal features in these rare non-stop aspirated sounds to distinguish them from their unaspirated counterparts. We propose to undertake this, first, by detecting the boundaries of frication and nasality in aspirated fricatives and aspirated nasal sounds, respectively. Following that, we plan to detect the presence or absence of aspiration in the aspirated non-stop consonants using acoustic evidence.

Aspiration in consonants is characterized by turbulent and voiceless noise [3, 32]. The noise-like properties of aspiration result in aperiodic glottal pulses with lower amplitude values. Presence of aspiration is also indicated by the excitation characteristics of the speech signal. Further, the vocal tract remains relatively open during the production of aspiration. These characteristics of aspiration can be captured using zero frequency filter (ZFF) [129], vocal tract constriction (VTC) [79], normalized autocorrelation peak strength (NAPS) [130] and strength of excitation (SoE) [129].

ZFF attenuates the higher frequencies, resulting in a better representation of the low frequency components, thus, facilitating the identification of aspiration. The excitation characteristics of aspiration can be seen in the zero-frequency filtered signal (ZFFS). The relatively open vocal tract configuration during aspiration can be captured by VTC information using ZFF. Here, a consonant, which assumes greater vocal tract constriction will result in higher VTC values; whereas, aspiration will result in lower VTC values. The aperiodic glottal pulses with lower amplitude values associated with aspiration, can be captured using normalized autocorrelation peak strength (NAPS) and strength of excitation (SoE) measures.

Using the aforementioned measures, in this study, we propose a methodology for automatic detection of aspiration in aspirated fricative /s^h/ and aspirated nasals /m^h/, /n^h/ and /$ŋ^h$/ in Rabha and Angami languages. While Rabha, a Tibeto-Burman language of the Bodo-Garo family, has the aspirated /s^h/, it does not phonemically contrast it with /s/ [36]. Angami, another Tibeto-Burman language, has phonemically contrasting /m/- /m^h/, /n/-/n^h/ and /ŋ/-/$ŋ^h$/ pairs in the language (see Table 4.2). Considering this, in this work, we explore the possibility of using source features in

these rare non-stop aspirated sounds to distinguish them from their unaspirated counterparts. We propose to undertake this, first, by detecting the boundaries of frication and nasality in aspirated fricatives and aspirated nasal sounds, respectively. Following that, we plan to detect the presence or absence of aspiration in the aspirated non-stop consonants using acoustic evidence [36, 43].

The rest of the chapter is structured in the following manner. Section 4.2 of the chapter discusses the properties of aspiration in stops, fricatives and nasals and provides justification for using acoustic features in automatic detection of aspiration. Section 4.3 discusses the acoustic features proposed in this work and Section 4.4 provides a description of the languages, linguistic databases and methodology. Section 4.5 describes the approach used for automatic detection of aspiration. Section 4.6 discusses the results of the study and finally, Section 4.7 summarizes the work.

4.2 Properties of Aspiration

The turbulent airflow associated with aspiration resembles the spectrum of a short [h] fricative before voicing of the vowel [3]. Detection of aspiration is relatively easy when it occurs after a stop consonant, as the burst associated with stops provide a reasonably clear boundary that separates the aspiration from the stop. However, distinguishing fricatives from aspiration becomes fairly difficult as both the frication and aspiration regions contain similar spectral characteristics. The properties of aspiration associated with stops are significantly different from the ones associated with fricatives and nasals. Unlike stops, an overlapping gesture of aspiration and frication or aspiration and nasal is observed in the fricatives and nasals. The articulatory and acoustic characteristics of aspiration in consonants with three manners of articulation, namely, stop, fricative and nasal have been discussed in Chapter 2.

4.3 Features for capturing aspiration

This section describes the features we propose to use in order to discriminate aspirated nasals and fricatives from their unaspirated counterparts. As mentioned earlier, Angami and Rabha are two languages that have aspirated fricatives and aspirated nasals in their consonant inventory. Hence, in the following sections, an automated method for detecting aspiration will be tested on data from

these two languages.

4.3.1 Aspirated Fricatives

In this study, the zero-frequency filtered signal (ZFFS) is analyzed and 4 features are exploited to capture the aspiration evidence in the aspirated consonants studied in this work. The 4 features, namely, strength of excitation (SoE), variance of successive epoch intervals (VSEI), vocal tract constriction (VTC) and normalized autocorrelation peak strength (NAPS), detect aspiration in the aspirated fricatives and aspirated nasals. While VTC, NAPS and SoE are used to distinguish between aspirated and unaspirated nasals, SoE and VSEI are used to distinguish between aspirated and unaspirated fricatives. These measures are extracted from a ZFF signal using the methodology summarized from (1) – (4), below [129].

- The speech signal is differenced

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

- Then, the signal $x[n]$ is passed through two cascaded ideal zero frequency digital resonators. The output of such a resonator is given by

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

where a_ks are the coefficients having value 4, -6, 4 and 1 and $p = 4$

$$z[n] = y[n] - \widehat{y}[n] \tag{4.3}$$

$$\widehat{y}[n] = \frac{1}{2N+1} \sum_{m=-N}^{N} y[n+m] \tag{4.4}$$

The output of the resonator is an exponentially growing signal due to the integration. To observe the excitation source signal, the trend from this signal needs to be removed by finding the mean of the quantity within a window of duration around the mean pitch period of the given signal. The trend removed signal $z[n]$ termed as 'zero-frequency filtered signal' (ZFFS), is achieved by (4.3) and

(4.4). Here, $\widehat{y}[n]$ represents the mean and $(2N+1)$ represents the mean pitch period over longer segments of speech.

4.3.1.1 Strength of Excitation (SoE)

SoE is extracted from the speech signal using the ZFF method. The slope of the ZFF signal $z[n]$ around the epochs gives a measure of SoE i.e $E[k]$ [129] and is computed as in (5), where, k=1,2,....,M and M represents the number of epochs. The SoE at k^{th} epoch e_k thus represents the relative amplitude of impulse-like excitation around that instant of significant excitation, namely the glottal closure instant (GCI).

$$E[k] = |z[e_k + 1] - z[e_k - 1]| \tag{4.5}$$

The SoE is expected to be larger for the nasals and vowels since they are periodic, voiced region with regular interval of zero crossings. The SoE is low in the voiceless regions, such as, the frication or aspiration regions. Therefore, when the the inherently voiced nasals are accompanied by inherently voiceless aspiration region, the SoE is expected to decrease during aspiration.

4.3.1.2 Variance of Successive Epoch Intervals (VSEI)

At the glottal closure instant, the significant excitation are said to be epoch, identified by the positive zero crossings of the ZFFS. Positive zero crossings are the points in the ZFFS where the signal crosses the zero reference line from negative to positive cycle. The zero crossings in the voiced region occur at regular intervals. Hence, the variance contour of the successive epoch intervals will be minimum in case of a sonorous, vowel like region, as seen in Figure 4.2 (d). In the voiceless region, the zero crossings occur at irregular intervals and therefore the variance will be large. Hence, in case of frication noise, the variance will be larger.

4.3.2 Aspirated Nasals

Unlike the voiceless, alveolar fricatives in Rabha, aspirated nasals in Angami are voiced. Therefore, the features used to capture vowel-nonvowel region in the aspirated voiced nasals of Angami will differ from the features used in aspirated voiceless fricatives in Rabha. In case of aspirated nasals, vocal tract configuration changes from nasal (high vocal tract constriction) followed by as-

piration (low vocal tract constriction). Hence, the VTC feature will be able to detect the initiation and termination of the nasal region. At the same time, NAPS ratio, calculated by taking the ratio of the ZFF signal and the raw signal, provides distinctive high values for the voiced regions. Hence, to determine the boundary between the nasal part and the aspiration of aspirated consonants, NAPS ratio is calculated. The motivation for using the VTC and NAPS ratio features are provided in the following sections.

4.3.2.1 Vocal Tract Constriction (VTC)

The constriction on the glottal vibration is affected by the obstruction of the airflow through the vocal tract system [131] and the effect is different for different sound categories [57, Figure 1]. In case of voiced speech, due to constriction in the vocal tract, the first formant decreases and the bandwidth of the first formant increases. The overall amplitude of the spectrum decreases with a relatively higher attenuation in the higher frequencies. As a result of that, relative amount of low frequency components increase in the constricted sounds, compared to other sounds [79].

The VTC evidence is measured with a cosine kernel value (V), which is directly proportional to the relative amount of constriction in the vocal tract. The method to compute the objective measure 'V' [79] is shown in (6) where $s[n]$ and $z[n]$ represent the speech signal and the corresponding ZFFS respectively between successive epochs. The unit-less cosine kernel value 'V' is a measure of similarity between $s[n]$ and $z[n]$.

$$V = \frac{< s[n], z[n] >}{||s[n]||||z[n]||} \tag{4.6}$$

The normalized values in the VTC evidence (V) are the highest for voiced bars due to the complete closure of the vocal tract. Contrarily, VTC values are the lowest for low vowels which are produced with a wide open vocal tract configuration [79, Figure 3]. In the case of nasals too, vocal tract is completely closed and the presence of nasal tract introduces a low first formant at around 250 Hz. This results in a high VTC value for nasals as well. Figure 4.1 shows the waveform and corresponding VTC values for a nasal sound followed by a low vowel. It can be seen that the VTC evidence has higher values in the nasal region and has lower values in the vowel region.

In case of aspirated nasals, high frequency components are introduced in the spectrum due to

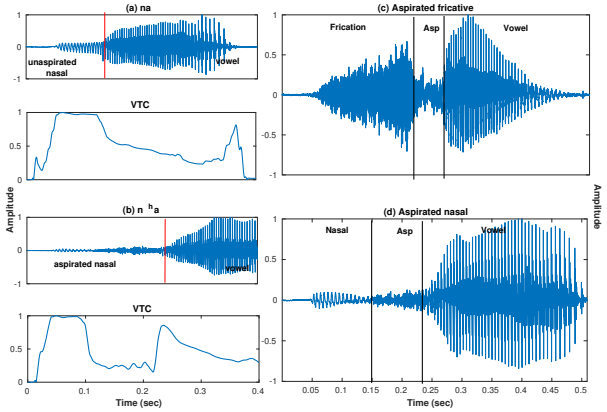

Figure 4.1: Illustration of waveform and VTC evidence for (a) unaspirated nasal (b) aspirated nasal and example of data annotation with (c) frication, aspiration and vowels marked for /s^{h}am/ (d) nasal, aspiration, vowel marked for /m^{h}a/

aspiration. Since VTC evidence is based on the relative amount of low frequency components, introduction of high frequency components in the spectrum reduces the VTC value. Therefore, the overall VTC value for the aspirated nasals will be lower, compared to that of the unaspirated nasals. We explore this nature of the VTC values to discriminate between the aspirated and unaspirated nasals. Figure 4.1 (a) and (b) show the waveform and the VTC evidence for an unaspirated and an aspirated nasal. It can be clearly seen that the VTC has lower values for the aspirated nasal compared to the unaspirated one.

4.3.2.2 Normalized Autocorrelation Peak Strength (NAPS)

A short-term autocorrelation is performed on the ZFFS. The value of the first peak (after the central peak) in the autocorrelation sequence is divided by the value of the central peak and the result is referred to as the normalized autocorrelation peak strength (NAPS). NAPS emphasizes the periodicity information of the signal [130].The pitch period will be random at the aspiration duration and will be uniform for the periodic part of the signal. Wherever the pitch period location is detected, the peak amplitude over the center value is computed for each frame, which, gives us the NAPS. NAPS is expected to give high evidence for voiced regions due to their periodic nature and low for the voiceless regions. The sinusoid like nature of ZFF is exploited and therefore the NAPS computed for the ZFFS would be high throughout the ZFFS of the samples containing a consonant-vowel portion.

The NAPS ratio ($\hat{R}$) is computed by the taking the ratio of the NAPS of the original speech sample and NAPS of the corresponding ZFFS of that sample for each frame as seen in (7).

$$\hat{R}[n] = \sum_{n=-N}^{N} \frac{R(s[n]))}{R(z[n])} \tag{4.7}$$

In (7), R is the NAPS and $s[n]$ is the raw speech signal, whereas, $z[n]$ is the ZFF transformed signal. NAPS is calculated between the successive epochs. During aspiration, NAPS will be low for the raw speech signal, due to its aperiodicity. On the other hand, NAPS will be high for the ZFF transformed signal as it induces periodicity in the signal. As a result, the NAPS ratio will be low for the aspirated or the fricative regions. However, for unaspirated nasals, the NAPS would be high for both the raw speech signal and the ZFF transformed signal, resulting in higher, near 1, NAPS ratio values.

4.4 Database Information

The Rabha speech database was collected from 44 native speakers of Rabha. There were 9 female and 13 male speakers from the Rongdani variety and 6 female and 16 male speakers from the Maitori variety. The participants could also speak Assamese and Hindi. For Angami, speech data of 21 individuals (12 male and 9 female speakers) were recorded reading a list of words. The speakers' age range was 20 to 50 years and have high school as the minimum education and thus could read the text provided for them to read.

The target words were repeated in three different environments: in isolation,in a fixed phrase, and in a sentence frame. The response was recorded with a Shure SM10 microphone attached to a Tascam DR 100 MKII solid–state audio recorder in .wav format. The speech samples were recorded at a sampling frequency of 44.1 kHz and were manually annotated with PRAAT [4]. All recordings were conducted in quiet environment. The speech tokens were marked for frication or nasal, aspiration and vowel regions, as seen Figure 4.1 (c) and (d), by careful listening and by visual inspection of the spectrogram and waveform.

4.4.0.1 Aspirated fricatives

1144 monosyllabic utterances (re-sampled to 16 kHz) of the voiceless alveolar fricative have been analyzed for this study. Participants for the study are selected from two varieties of the language, namely, Rongdani and Maitori. The participants read out a list of words with tokens containing the voiceless alveolar fricatives shown in Table 4.1. As seen in the table, aspiration in the fricatives in Rabha is noticed only in the low vowels. The voiceless, alveolar fricative is unaspirated in the other vowel contexts.

Table 4.1: Monosyllabic word list of the form /CV/ and /CVC/ (C: consonant,V: vowel)

Word	Gloss
sʰa	eat
sʰak	leaf
sʰam	wait
sʰam	grass
seŋ	waist

Word	Gloss
si	blood
siŋ	ask
so	mosquito, burn, rot
soŋ	village
su	wound

4.4.0.2 Aspirated nasals

Speech data of 21 individuals were recorded reading a list of words. The tokens with nasals (resampled to 8 kHz) were selected for analysis. The Angami database used in this study has 285 utterances with 155 aspirated nasals and 130 unaspirated nasals. The monosyllabic nasal utterances have 7 minimal pairs, contrasting in aspiration as shown in Table 4.2. The collected database is from an under-resourced language, however, the size of the database is sufficient to be used in a knowledge based classification analysis.

Table 4.2: Angami wordlist with /CV/ syllables

Word	Gloss
mʰa	something to work
ma	people/price
mʰi	eye
mi	light/fire
mʰo	above
mo	cow mooing
ne	push
ni,nii	earring,happy

Word	Gloss
nʰa	plants
na	blackseed/budding
ŋʰa	messy
ŋa	ripe/crazy/tired
nʰo	plastic cover
no	child/suffocated/you
nʰu	something to carry
nu	breastfeeding

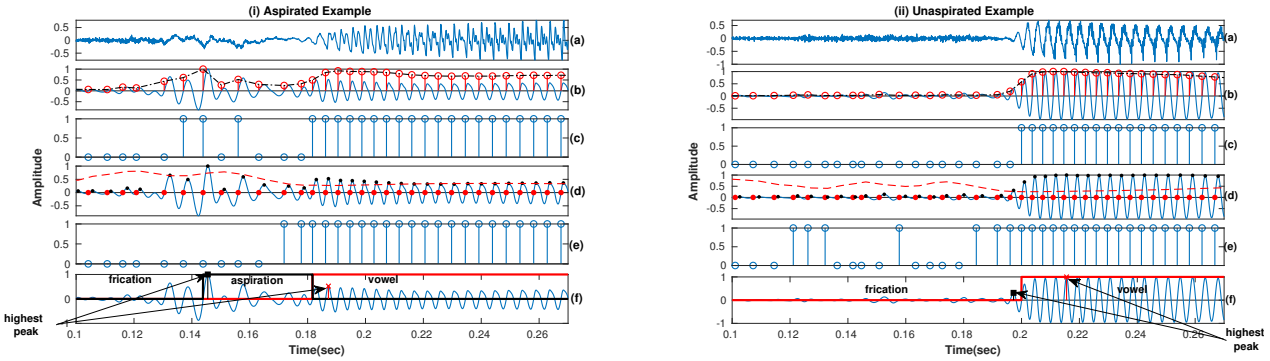

Figure 4.2: Illustration of the proposed method for (i) Aspirated fricative /s^{h}a/ and (ii) Unaspirated fricative /si/. (a) Speech signal sampled at 16 kHz (b) Excitation strength plot (dashed black line) of the ZFF signal (c) VLR decision plot with respect to the excitation strength (d) Variance plot (dashed red line) of epoch intervals (red dots); black dots indicate the positive peaks between successive epochs (e) VLR decision plot with respect to '(d)' contour (f) Vowel and non-vowel like region; red box indicates vowel region; black square and red cross indicate the maximum peaks in NVLR and VLR, respectively.

4.5 Automated approach

4.5.1 Aspirated Fricatives

The measurements of the acoustic study in the aspirated fricatives suggest the presence of more information in the low frequency region. There is high energy in the low frequency region for aspiration as compared to frication region (see Figure 2.2). Using this information, a method is proposed to detect the presence or absence of aspiration in fricatives in Rabha monosyllables.

In the production of the aspirated fricatives, an unvoiced excitation is present along with the excitation source signal. The change in the excitation characteristics can be used for the detection of presence or absence of aspiration in fricatives. For this, the speech signal needs to be processed to extract excitation source information. As the ZFF method provides reliable information about the excitation source characteristics [129, 132], the proposed method in this chapter uses ZFFS for automatic detection of existence of aspiration.

The excitation source information becomes more pronounced in the ZFF signal and thereby, various characteristics of the excitation source signal can be observed. The ZFFS is periodic in the vowel region due to the glottal vibration responsible for the production of vowels, while it is aperiodic in the fricative region due to the random noise like excitation during the production of fricatives. Also, the large amplitude of ZFFS indicates the strength of excitation associated with

the production of a particular segment of speech signal. The ZFF method attenuates high frequency signal and emphasizes the F0 region. It captures the low frequency components of aspiration and results in a large impulse like excitation in that region. Therefore, these impulse like characteristics can be attributed to the existence of aspiration in the fricative. Taking this assumption as the basis, a new approach is suggested for detecting aspiration in fricatives, as described below.

(i) Find positive zero crossings of the ZFFS (shown by a red * in Figure 4.2(d))

(ii) Detect the peak within two successive zero-crossings (shown by the dotted line in the Figure 4.2(d))

(iii) Estimate the location of the largest peak in the signal

(iv) If the largest peak occurs in the frication region, then the corresponding fricative is considered an aspirated fricative, else, as an unaspirated fricative.

A-priori knowledge of vowel and fricative regions is assumed for the above steps. However, in an automatic method, *a-priori* knowledge of the segment boundaries may not be available. Therefore, the ZFFS is used for broad classification of the given monosyllables into vowel and fricative regions. The interval between successive positive zero crossings will roughly correspond to pitch period in case of a vowel like region (VLR). The variance contour of the successive positive zero crossing intervals, i.e., the epoch intervals will be minimum, in case of vowel like region and it will be large, in case of the fricative region, as seen in Figure 4.2(e). Moreover, a higher SoE value in the ZFF signal indicates a VLR. These two features are used to distinguish the speech sample into two parts, namely, VLR and non-vowel like region (NVLR). The NVLR can be consonant or fricative regions. The steps involved in distinguishing the speech sample into VLR and NVLR areas are detailed as follows.

(i) Step I : Determine the strength of excitation (say E) of the ZFFS.

- if $E(i) > (1/2) * max(E)$, then consider the region between i^{th} and $(i+1)^{th}$ epoch locations as a vowel region. Here, max(E) represents the maximum excitation strength of the signal.

(ii) Step II : Calculate the epoch interval values for all the epoch locations and compute the variance for the successive epoch intervals using overlapping frames.

- If $v(i) < mean(\mathbf{V})$, then consider the region between i^{th} and $(i+1)^{th}$ epoch locations as the vowel region. Here, $\mathbf{V}$ represents the vector containing the variance values $v(i)$ of all the epoch intervals.

(iii) Combine decisions from Step I and Step II and find the common region that satisfies both the conditions. These regions are marked as VLR and the remaining regions as NVLR.

Table 4.3: Trade-off with respect to strength of excitation (E); Accuracies are in terms of %

Threshold	Accuracy(%)(MA)	Accuracy(%)(RA)
0.3	54	49
0.4	66	59
0.5	77	76
0.6	91	88
0.7	95	96
0.8	99	99

Table 4.3 shows the trade-off with reference to the threshold value chosen for the vowel like region, based on the strength of excitation for Maitori aspirated (MA) and Rongdani aspirated (RA) fricatives. To avoid a critical threshold, a generalized threshold is chosen experimentally, i.e. (1/2), of the maximum amplitude of E. Values above and below the chosen threshold give more spurious values due to the misclassification of vowel onset points. Eventually, these are expected to affect the classification accuracy. Even if a threshold value greater than 0.5 gives better accuracy, on analyzing the results, more spurious values were observed. Hence, to avoid that, 0.5 is taken as the threshold for the knowledge based classification method for vowel like region detection. Ideally, the presence of the largest peak in the consonant region indicates the presence of aspiration. However, in some cases, the peak may appear at the onset of the vowel region. Hence, an algorithm is proposed, as stated below, to hypothesize the presence or absence of aspiration in the fricative region. Based on this, a decision is made whether the given utterance is aspirated or not.

(i) Find maximum peak value of NVLR (say p) and maximum peak of VLR (say q)

(ii) Compare p and q

(a) if $p > q$, then consider the non-vowel region as an aspirated fricative

(b) else if $p < q$ and $p > (q/2)$, consider the NVLR as an aspirated fricative

(c) else, consider as unaspirated fricative

4.5.2 Aspirated Nasals

In order to automatically detect aspiration and to distinguish an aspirated nasal from an unaspirated one, the speech samples were re-sampled to 8 kHz and silence regions were removed. Short term energy (STE) of the speech samples was extracted and a very low threshold value is set to identify the silence region. Further, First Order Gaussian Differentiator (FOGD) is applied to further segment the samples into vowel and consonant regions. VTC, SoE and NAPS ratios are considered to identify aspiration in nasals, and based on the presence or absence of aspiration in the nasals, a decision is made whether the nasals are aspirated or not.

Individually, VTC, SoE and NAPS ratio, may not be able to detect aspiration on nasal consonants. Hence, all three features are combined and used for aspiration detection in Angami nasals in the non-silence regions. The key steps involved in the detection of aspiration in Angami nasals are summarized as follows.

(i) Compute the ZFFS of the signal

(ii) Analyze VTC from the ZFFS

(iii) Extract SoE at each epoch from the ZFFS

(iv) Find the NAPS for the raw speech signal and the NAPS for its corresponding ZFFS

(v) Compute the 'NAPS Ratio' between the NAPS of the raw speech signal and the NAPS for its corresponding ZFFS

(vi) Normalize the three evidences

(vii) Combine all the three evidences (VTC, SoE, NAPS Ratio) by taking their sum

(viii) The non-vowel regions where the combined sum follows a high to low transition value are considered aspirated nasals , else unaspirated nasals.

As discussed in case of the aspirated fricatives, *a-priori* knowledge of the consonant and vowel boundaries is usually not available for the automatic method. At the same time, the methodology applied for aspirated fricatives is not applicable for the aspirated nasals as the nasals are also voiced, periodic portions, resembling the vowel parts of the consonant-vowel combination. This makes it difficult for the automatic method to distinguish between a nasal and a vowel. However, it is seen that there is a significant energy difference between a vowel and a nasal, as the nasal membranes absorb energy. The First Order Gaussian Differentiator (FOGD) captures such changes in energy in the signal. Hence, for the aspirated nasals, FOGD is applied on the Short Term Energy (STE) of the non-silence regions of the speech. This helps us distinguish the nasal periods in the signal from the vocal periods. Hence, the complete steps to identify aspiration, nasal and vowel boundaries are provided as follows.

(i) Find the STE of the speech signal

(ii) Set a very low threshold value (*approx.zero*), the regions falling below the threshold are considered silence

(iii) All the observations are made within the non-silence regions

(iv) Find the FOGD of the STE of that region

(v) The regions with higher FOGD values are considered vowels, ones with lower values are considered non-vowels

(vi) Change in energy from non-vowel to vowel is captured and the location where the slope changes from low to high is considered as the boundary between a vowel and a consonant

(vii) Observe the change in slope in the combined evidence plot ($VTC + SoE + NAPSratio$) in the detected consonant region

(viii) If the change of slope occurs from high to low, consider it an aspirated nasal, else consider it an unaspirated nasals.

Table 4.4: Performance analysis of the different subgroups of Rabha (performances are done independently for each of the cases)

Groups	Accuracy in percentage
Maitori Aspirated Fricatives	77 %
Maitori Unaspirated Fricatives	57 %
Rongdani Aspirated Fricatives	76%
Rongdani Unaspirated Fricatives	63 %

4.6 Results and Discussion

4.6.1 Aspirated Fricatives

In case of the aspirated fricatives, the ZFFS can convincingly indicate the presence or absence of aspiration. As seen in Figure 4.2(i) (b), in the automated approach, it is observed that the existence of high positive peak of the ZFFS in the consonant region is associated with the presence of aspiration in a fricative. The high peaks in the ZFFS occur due to the significant excitation strength at the epochs in the aspiration part of the consonant region.

Also, the ZFF based features, such as, strength of excitation and variance of epoch intervals can determine the vowel and consonant regions of the aspirated fricatives as observed in Figure 4.2(i) (f).As discussed earlier in 4.5.1, regions with high excitation strength and low variance contour of the epoch interval are considered vowel regions, marked by a red line in Figure 4.2(i)(f), and the rest as fricative regions. If the largest amplitude of the positive peak in the NVLR (p) is greater than that of the VLR (q), the fricative region is assumed to have aspiration and the region from the largest peak location in NVLR to vowel onset point is marked as aspiration region, marked by a black line in Figure 4.2(i)(f).

In the cases where the largest amplitude of the positive peak is observed in the VLR (q), the fricative regions where twice the value of p is greater than q, are considered aspirated.

4.6.1.1 Performance evaluation

The primary objective of the chapter is to detect the aspiration in the aspirated fricatives. However, in order to get a clear picture of the performance of the automated analysis, the ZFF method is performed on the speech signal for both the Maitori and Rongdani (two subgroups of Rabha) databases, once with aspiration and once with the unaspirated case. As compared

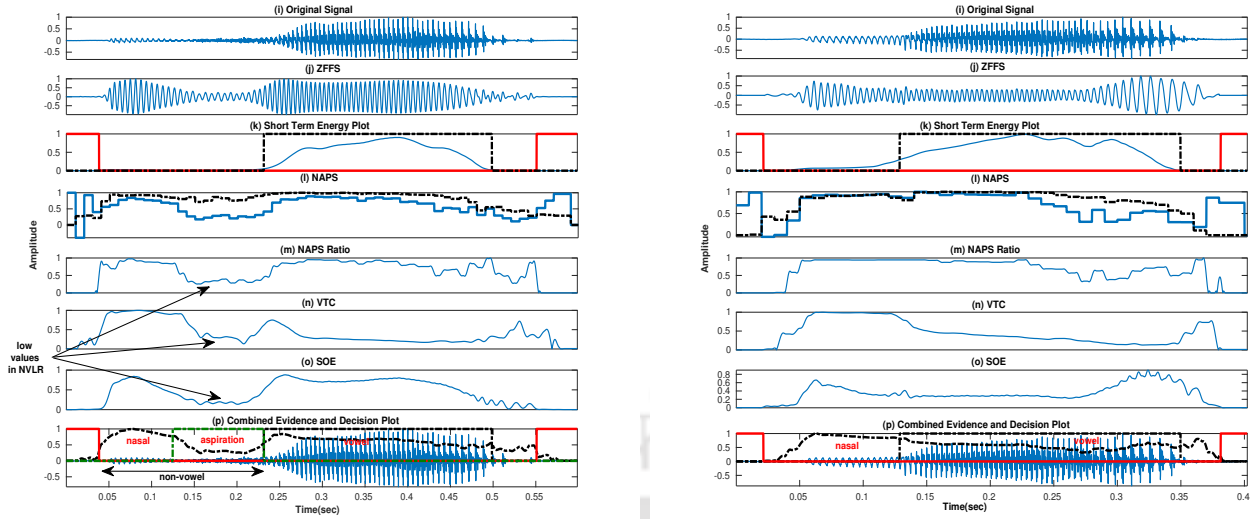

Figure 4.3: Illustration of the proposed method for (a) aspirated nasal /m^{h}a/ (left) & (b) unaspirated nasal /na/ (right) (i) Speech signal sampled at 8 kHz (j) ZFF signal (k) strength of excitation(SoE) (l) Short Term Energy (red line showing silence region, black dotted line showing vowel region)(m) NAPS (blue line for speech signal,black dotted line for corresponding ZFFS) (m) vocal tract constriction (VTC) evidence, (o) NAPS ratio , (p) Combined evidence (black dotted curve) with red line showing silence region, green dotted line showing aspiration region, black dotted line showing vowel region

to aspirated ones, the excitation strength computed at the GCIs is small in the NVLR in the unaspirated ones, as shown by the dashed black line in the Figure 4.2 (ii)(b). In most of the cases of the aspirated group, it is seen that the peak obtained from ZFFS is maximum at the consonant portion as compared to the vowel portion and hence considered as aspiration. In few cases, the maximum peak value is obtained at the vowel region. However, in such cases, evidence of aspiration is still present since the aspiration peak is quite large (more than 50% of the maximum peak value in the vowel region), so, even these consonant are said to have aspiration region. This trend is not observed for the databases with unaspirated fricatives, i.e., although some peaks may occur at the frication part, it is almost negligible compared to the large peak value in the vowel region, as seen in the Figure 4.2 (ii)(f). Hence, in such cases, no boundary is defined for the aspiration region.

The performance of the automatic detection of aspiration of Rabha language is evaluated on the enhanced database. The classification into aspirated or unaspirated token prepared by the expert phoneticians were considered as the ground truth. The analysis is carried out for the different subgroups in isolation (as mentioned in Table 4.4). The aspirated groups have shown an accuracy of average 77% (better than chance), while the unaspirated groups have showed an accuracy of

average 60%. When the failure cases are analyzed for both the groups, one of the three things is observed - (i) peak values fall below the threshold, (ii) VLR and NVLR are not segmented properly and (iii) the speech segment is not of 'CV/CVC' form, i.e, the monosyllable is preceded by some utterances/noise. So, additional features may be required along with enhanced threshold values for better performance of the automated method.

4.6.2 Aspirated Nasals

Though the ZFF method is expected to give excitation information regarding aspiration in the signal, the nasals being voiced followed by vowels, the sinusoidal nature is enhanced in the entire signal. Illustration of the ZFFS (Figure 4.3) (j) shows that there is low amplitude during the aspiration phase. A comparatively high amplitude of the ZFFS is observed in the vowel and the nasal part of the consonant region in Figure 4.3 (j) of the aspirated example. This is because the ZFF captures the low frequency information and attenuates the higher frequencies. Hence the higher frequencies in the noise like aspiration region gets attenuated and a comparatively low amplitude of the ZFFS is observed in that region. The proposed method for vowel-nonvowel segmentation in Angami nasals is evaluated and silence region and vowel region is obtained, as seen in Figure 4.3 (k).

In Figure 4.3 (n), the trend of the VTC evidence is high in nasal and low in vowel. As discussed earlier, the constriction is high in nasals compared to wider vocal tract opening in nasals. When nasals are accompanied by aspiration, the introduction of high frequency components reduces the VTC value in the aspiration part of the consonant region. During the production of aspiration, the vocal tract is released, as a result, the constriction evidence further lowers before the utterance of the following vowel. In higher vowels, this trend might not always occur due to moderate constriction as compared to low vowels. Hence, the difference between the VTC evidence of the aspiration and high vowels may not be well distinct. In such case, requirement of some other evidence is necessary to distinguish the aspiration region. For that purpose, we move forward with the SoE and NAPS ratio. In Figure 4.3 (l), the NAPS for the original raw signal is represented by the blue line while the black dotted line represents the NAPS of the corresponding ZFFS. The NAPS ratio is observed to be the lowest in the aspiration region (See Figure 4.3(m)). It is observed that the NAPS ratio

Table 4.5: Performance analysis of the Angami nasals (performances are done independently for each of the cases)

Groups	Accuracy in percentage
Aspirated Nasals	75 %
Unaspirated Nasals	85%

gives some abrupt values due to the random NAPS value (for speech and their corresponding ZFFS) in the beginning and end of the speech samples. Hence, NAPS ratio is computed within the non-silence region to avoid unnecessary confusion. The amplitude observed in Figure 4.3 (o) indicates the relative strength of excitation around the epoch and is high for the nasals and vowels, leaving out aspiration with lower amplitude.

Thus, the combination of all the three evidence, as seen in Figure 4.3 (p), within the non- silence region is used to clarify the distinction of the aspiration region. The sharp transition in slope from high to low evidence in the consonant region marks the nasal to aspiration transition. This trend is captured and analyzed to mark the aspiration region until the vowel onset point. Thus, we get silence, nasal, aspiration and vowel region for each speech sample as shown in Figure 4.3 (p). For the unaspirated nasal, there is no aspiration and the automated method, therefore, classifies the entire speech signal into silence, nasal and vowel regions with no aspiration region detected (shown in Figure 4.3 (p)).

4.6.2.1 Performance evaluation

Considering the ground truth based on the perception and acoustic analysis by expert phoneticians, the performance of the automatic detection of aspiration of Angami language is evaluated on the enhanced database. The analysis carried out in isolation (as mentioned in Table 4.5) shows an accuracy of 75% for the aspirated groups while 85% for the unaspirated groups. Examples with bi-syllables, nasals cut-off during the data segmentation in PRAAT and no prominent aspiration has been taken care of during the automatic analysis. During the error analysis, it is found that in some cases, out of all the three evidences(i.e VTC, SoE, NAPS ratio), only 1 or 2 feature has shown the aspiration evidence. As a result, spurious cases are observed during the analysis of the combined evidence. Apart from that, it is observed that (i) VLR and NVLR not segmented properly and (ii) aspiration could not be captured properly. One of the reasons for the aspiration not

being captured might be due to the high vowels when the vocal tract production phenomenon of aspiration and high vowel merges together. In few cases, the vowel region is detected much before the vowel production, hence the aspiration evidence could not be captured. The aspiration can be affected by the following vowels as well as the speakers. All these factors are expected to affect the classification accuracy.

4.7 Summary and Conclusion

This chapter characterizes the property of aspiration in nasals and fricatives in two languages, Rabha and Angami. While aspiration in stops is typo-logically widely prevalent, aspiration in nasals and fricatives is rare and characterized by idiosyncratic acoustic characteristics. Aspiration in stops is characterized by a period of voicelessness following the release of the stop, and a strong burst at the release. On the other hand, aspiration in fricatives and nasals is characterized by lower vocal tract constriction, aperiodic glottal pulses with low amplitude and expressed in measures such as, VTC, NAPS ratio, SoE and VSEI. These features can distinguish aspirated consonants from their unaspirated counterparts. In this work, the VTC, NAPS ratio, SoE and VSEI features are used for classifying aspirated from unaspirated nasals and fricatives in Angami and Rabha. These features are expected to improve the performance of the PRS (Phone recognition system) in Angami, which will be demonstrated in the succeeding chapter work.

In the next chapter, different binary classifiers are combined for the classification of two-way contrast of fricatives in Rabha using temporal information. The temporal features along-with the features explored in the current chapter for the Rabha fricatives will be further explored for automatic detection of aspirated fricatives for Korean, a widely-studied language that is non-related to Rabha, having aspiration contrast in fricatives. This would demonstrate whether the explored features are language independent or not.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

5

Acoustic phonetic correlates in detecting aspiration in fricatives of Rabha and cross-linguistic comparisons with Korean

Contents

5.1 Introduction

Aspiration in fricatives is a relatively rare phenomenon in languages, and only a small number of languages report aspirated fricatives as distinctive phonemes in their languages [14, 29, 34, 43, 51]. Apart from that, automatic detection of aspiration in aspirated fricatives is deemed a challenging task [28, 36, 43]. Even in the case of humans, the contrast between aspirated fricatives and unaspirated fricatives may not be perceived easily [28, 133]. It has been observed that aspirated fricatives occur in Rabha and Korean languages. In the previous chapter, automatic detection of the aspiration in Rabha aspirated fricatives was attempted using source features [43]. In the current work, apart from the source features, temporal features will also be explored for automatic classification of aspirated fricatives. Additionally, the classification method will also be evaluated on data from Korean language, which also has aspiration contrast in fricatives [34].

While several works have reported the acoustic differences between aspirated and unaspirated fricatives [36,41,44,48,89], only a few have attempted to automatically detect aspiration in aspirated fricatives [36,43]. One approach adopted in a previous work used two acoustic features: the strength of excitation (SoE) and the variance of successive epoch intervals (VSEI), in detecting aspirated fricatives in Rabha, using a knowledge-based classifier and yielded an accuracy of 76.5%. In the same work, when the two acoustic features were appended to the 39 dimensional MFCC features and fed into an SVM classifier using a linear kernel, two-class classification of Rabha fricatives yielded an accuracy of 60%, which was a 2% increase from the baseline using 39 dimensional MFCC features [43]. In another work [48], the best classification accuracy of 79.74% was reported for the Korean alveolar fricative class in bi-gram modeling using HTK toolkit and F0 cues. However, the confusion between the two types of fricatives, namely, aspirated and unaspirated, is not mentioned in the work.

The work in the previous chapter was limited in various ways. Firstly, it was not clear whether the proposed acoustic features to detect aspiration in aspirated fricatives are language-independent or not. Secondly, it did not take temporal features into consideration in the detection and classification of aspirated and unaspirated fricatives. Thirdly, in the previous work [43], the tuning parameters in the SVM were explored only in the linear domain using a linear kernel, and the effectiveness of the classifier using non-linear decision boundaries was not attempted. In this work, we extend

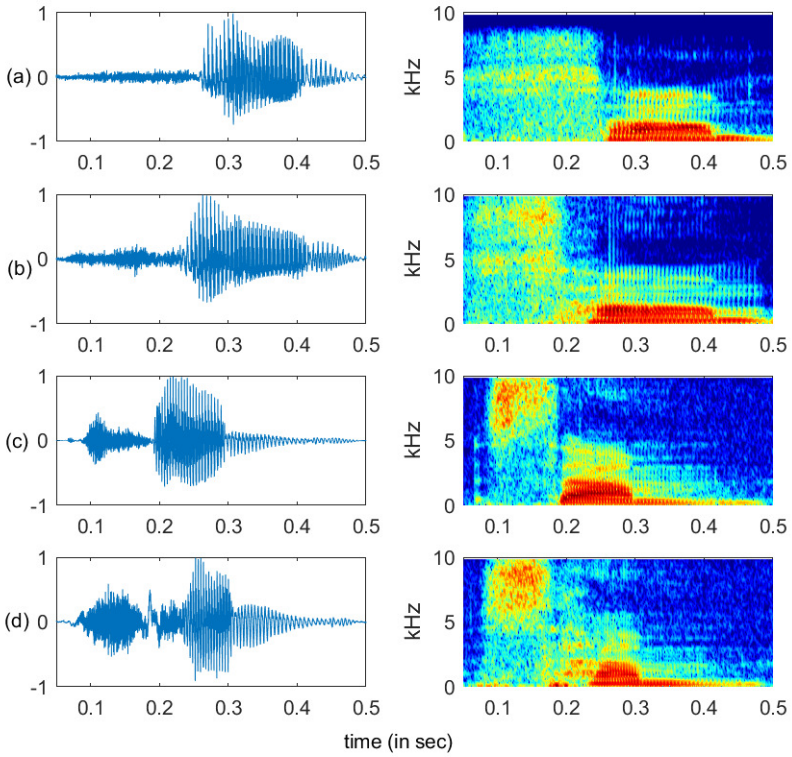

Figure 5.1: Time domain waveform and spectrogram of (a) Rabha unaspirated fricative /sam/ (b) Rabha aspirated fricative /s^{h}am/ (c) Korean unaspirated fricative /sam/ (d) Korean aspirated fricative /s^{h}am/

our previous work and continue the experiments by exploring other kernels in the SVM classifier. Moreover, the SVM parameters are searched empirically through a trial-and-error approach, which results in a higher computational burden for the constrained optimization programming [134, 135]. To overcome this drawback, aspirated-unaspirated classification is performed using a GMM based back-end. The over-fitting problem can be addressed by using a Random Forest classifier (RF), which is an ensemble learning model. SVM, GMM, and RF are antithetical to each other in the sense that SVM, RF are discriminative and supervised models, whereas GMM is a generative and unsupervised model. The discriminative model learns the decision boundary between the classes without caring about how the data was generated. On the other hand, the generative model learns the actual distribution of the classes and models how the data was generated. The random forest works on the principle that from a randomly selected subset of the training set, a set of decision trees is created. Based on the majority of votes from each of the decision trees, the final class of the test object is predicted. RF has lesser variance but is more complex and time-consuming than the decision trees. Thus, as comparative methods, different classifiers have been explored to evaluate how the proposed features distinguish between the aspirated and unaspirated fricatives.

One way of dealing with such antithetic models is to combine them into a hybrid model. The fusion of the classifiers compensates for the drawbacks of the individual classifier into the hybrid system. SVM hybrid classifiers have been explored for speech recognition [136], speaker identification [137, 138], dialect identification [139], etc. An SVM-GMM-RF hybrid classifier for classifying aspirated and unaspirated fricatives is applied in this work. This methodology is adopted by conducting the classification tasks based on the weighted scores obtained from the three models. The weight for this output is set high if any classifier gives far superior performance, and the others are set low. The best weight set for the individual GMM, SVM and RF classifiers is selected for the score level hybrid system to obtain the best classification performance.

5.1.1 Challenges and goals of the study

In order to ascertain the role of acoustic and temporal features in distinguishing between aspirated and unaspirated Korean alveolar fricatives, a series of perception experiments were conducted [34, 62]. The perception experiments showed that aspiration duration played a major role in

distinguishing Korean alveolar fricatives. As the aspiration duration decreased, perception shifted from aspirated to tense fricatives. It was also noted that fricative aspiration contrasts are difficult to perceptually distinguish than aspiration contrasts in stops or affricates [13]. Previous studies have reported varied duration for aspiration in fricatives. While some studies report that most unaspirated /s/ (or /s*/) tokens have zero or almost zero aspiration [34,62,63], other studies report aspiration duration in tense fricatives to be as long as 17 ms, with smaller differences between /s^h/ and /s/ [46]. It was also found that identification accuracy of the fricative contrast in Korean was 100% in the /a/ context, which reduced to about 80% in the /i/ context [63]. Hence, it is clear that measuring and detecting aspiration in voiceless fricatives is challenging and maybe subjective. Mismatch of acoustic and perceptual cues in distinguishing aspirated and unaspirated fricatives suggests an exploration of more robust features to separate the two categories of fricatives [28]. Therefore, in this chapter, we explore the performance of features such as relative duration, the strength of excitation (SoE), and variance of successive epoch intervals (VSEI) in classifying aspirated and unaspirated fricatives in the Rabha and Korean language. Although SoE and VSEI were used in an SVM classifier to classify Rabha aspirated and unaspirated fricatives, in the current study, we expand the work by adding the relative duration feature on GMM, RF classifiers. Apart from that, we also investigate the effectiveness of our system by classifying Korean aspirated and unaspirated fricatives [140]. The rest of the work is organized as follows: Section 5.2 describes the method for deriving the proposed evidence, respectively. Section 5.3 shows results and analysis of the evidence in various classifiers and languages. Section 5.4 summarizes the work, provides conclusions, and mentions the future scope.

5.2 Methodology

5.2.1 Feature extraction

The 512 Rabha tokens and 639 Korean tokens, described in Chapter 3, were used for training and testing the models. The raw signal, re-sampled to 16 kHz, is fed as input and is processed through the Zero-frequency-filter(ZFF) to get the Zero-frequency-filtered signal (ZFFS) [43, 129]. The features, namely, the strength of excitation (SoE) and the variance of successive epoch intervals (VSEI), are extracted using the feature extraction method explained in a previous work [43]. One of

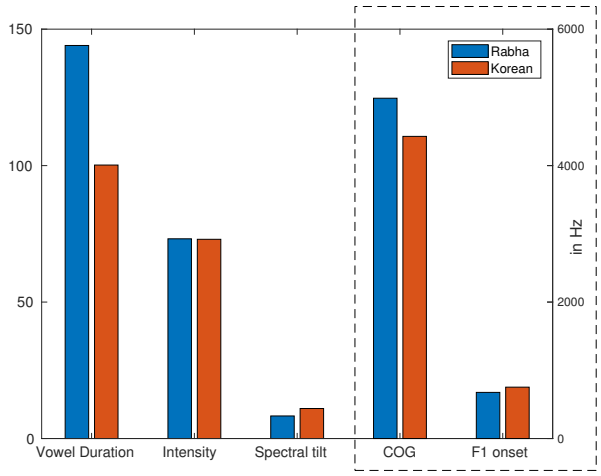

Figure 5.2: Temporal and spectral measurements of Rabha and Korean aspirated fricatives. Duration is measured in msec, intensity and spectral tilt is measured in dB.

the measures used to detect the aspiration characteristics is the strength of excitation (SoE), which is represented by the relative amplitude around the glottal closure instants of the ZFFS. Another measure, the variance of successive epoch intervals (VSEI), is extracted by calculating the average of the squared differences of the consecutive epoch intervals obtained from the ZFFS. In order to detect the aspiration region in the speech sample's frication segment, the speech sample needs to be segmented into vowel and non-vowel regions. The SoE and VSEI features are extracted, and a threshold is set to mark the vowel and non-vowel regions. The largest peak location is searched locally in both the regions and the speech segment from the onset of the largest peak location in the non-vowel region to the onset of the vowel region is marked as the aspiration region. The aspirated fricatives are expected to have this aspiration region, whereas the absence of this aspiration region would indicate that the fricative is unaspirated. Based on the apriori knowledge of vowel and non-vowel regions, a classifier is modeled, which decides whether the corresponding speech sample belongs to aspirated fricative or not.

Apart from the proposed features, Mel Frequency Cepstral Coefficients (MFCCs or M, in short) are also extracted. A frame-size of 20 ms with a frameshift of 10 ms for each speech signal is considered for our experiment. For each frame, 13 MFCCs, their delta and delta-delta coefficients are extracted, constituting the 39-dimensional MFCC features. The 39-dimensional MFCC features capture the vocal tract information. The source features SoE and VSEI are concatenated to the

39-dimensional MFCC features making it 41-dimensions (abbreviated as M-S-V in this chapter). During the feature extraction of each speech utterance, the SoE values are obtained at every sample point. Then for each frame, the mean of the sample values within that frame is computed. E.g., if the length of the k^{th} speech utterance has X_k sample points, then the number of frames would be

$$Y_k = \frac{X_k - p}{q} - 1 \tag{5.1}$$

where p, q are frame-size and frameshift, respectively. A speech utterance with the SoE feature would give $Y \times 1$ dimensional vector, while that with the MFCCs would give $Y \times 39$ dimensional features, one frame per row. When the SoE and VSEI features are appended to the MFCC features, we have $Y_k \times 41$ feature dimensions per speech utterance. The temporal acoustic parameter, the relative duration, is calculated once for the frication segment per speech utterance. The same value of the duration feature per speech utterance is repeated Y times and appended to the 41 dimension features (abbreviated as M-S-V-D in this chapter). The relative duration feature (r) has been computed by the following Equation 5.2. In other words, the relative duration is the ratio of the duration of the target phone and the sum of the total duration of the N phones for each speech sample.

$$r[k] = \frac{d[k]}{\sum_{n=1}^{N} d[n]} \tag{5.2}$$

where $d[k]$ is the duration of the k^{th} phone and $n = 1, ..., k, ..N$ represents n number of phones.

5.2.2 Feature analysis

To analyze the effects of the different features, the visual distribution maps are presented for the MFCC features and a combination of the proposed features appended to the MFCC features. The MFCC features consist of 39 dimensions, while the M-S-V-D comprises 42 dimensions, which include the 39-dimensional MFCC, SoE, VSEI, and relative duration features. The labels for the train-test set were set based on the ground truth annotation. The speech utterances were manually tagged by expert phoneticians as aspirated/ unaspirated based on the auditory and spectrographic evidence. One of the common dimensionality reduction methods is the t-distributed stochastic neighbor embedding (t-SNE), which helps visualize higher-dimensional data in low-dimensional

space. For the feature analysis, the Korean database is used to illustrate the visual distributions of the two-class, i.e., aspirated and unaspirated fricatives, using a t-SNE plot.

5.2.3 Classification

For the training-testing protocol, a four-fold validation scheme is adopted. In other words, the database is divided into four equal parts, and three-fourths of the database is fed as a training set to the chosen classifiers, and the omitted part of the database is used for testing. The classifiers are designed independent of the speakers and a mutually exclusive train and test set. In order to remove bias introduced by humans into classification results, the training and testing data-set is randomly selected. The classification algorithm is validated four consecutive times, where every training and testing set is independent of the other sets, and each fold is made the testing set in turn. The 39-dimensional MFCC feature is considered as the baseline method, and the performance for two-way-contrast fricatives, i.e., aspirated and unaspirated fricatives, are evaluated for baseline ($39MFCCs$) and proposed methods ($39MFCCs + SoE + VSEI + RDur$) using the chosen binary classifier for both the languages, i.e., Rabha and Korean.

For the SVM classifier, pairs of tuning parameters (C,γ) are evaluated with cost and gamma values being selected within the range 0 to 1, and $exp(-2)$ to $exp(2)$ with a step of 0.1, respectively and the one with the best average cross-validation accuracy is chosen. Different kernels are explored, and out of all the kernels, the radial basis function works best for our non-linear database. The GMM is parametrized to the different features, and the number of mixtures for the GMM predictions is tuned within a range from 1 to 25. For the random forest predictions, the two parameters, viz, the number of trees, and maximum features per bag, are tuned within a range of 50-300 and 1-10, respectively, using a grid search in this study. The tuning parameters for all the three classifiers vary for each combination of the train and test sets, and the optimal hyperparameter values are chosen, which gives the best performance. Finally, the average of all the accuracy values obtained for all the train-test sets has been considered for computing the overall accuracy for each of the classifiers.

Along-with the above-mentioned non-nested cross-validation (CV), a nested cross-validation strategy is also evaluated in the Rabha and Korean data-set. The training set is split into training

and validation sets, and the optimal hyper-parameter is chosen using a grid-search algorithm in the inner loop. To generalize to unseen data, we measure the ability of the model to generalize on a held-out test set by classifying a held-out test data-set using the optimal value of the hyper-parameters in the outer loop [141, 142].

5.2.4 Hybrid Classifier

Combining the discriminative model and the generative model can improve the performance of a speech recognition system [136, 137, 139]. This chapter presents a new method to combine both the discriminative and generative models to use the two models. In the proposed GMM-SVM hybrid classification, the individual SVM and Bayes classifier using the GMM are constructed using each feature separately, i.e., MFCCs, SoE, VSEI, and relative duration. An SVM model is built using the 39-dimensional MFCC features; similarly, separate SVM models are built by training with SoE, VSEI, and relative duration features. This results in 4 sets of SVM classifiers. The individual features are fused, and the SVM models are trained with the combinations of these features. The GMM models and the random forest models are also built similarly. The probability estimates that are produced by the individual classifiers are then set with optimal weights.

During the training phase, all three classifiers are trained and tested using the same data to obtain the probability estimates for the two classes. The performance is calculated separately for each classifier. The GMM and SVM classifiers are combined at the score level. The scores from both the classifiers for each sample is considered, and the combined score (S_c) is calculated in the following way:

$$S_c = \alpha S_g + (1 - \alpha) S_s \tag{5.3}$$

where S_g, S_s denotes the scores obtained from GMM and SVM classifiers, respectively. Along with these two existing classifiers, RF is introduced into the hybrid classification approach, and the overall performance is examined. The scores of the random forest classifier and the combined SVM-GMM classifier are fused to build a hybrid classifier.

$$S_h = \alpha S_g + \beta S_s + \tau S_r \tag{5.4}$$

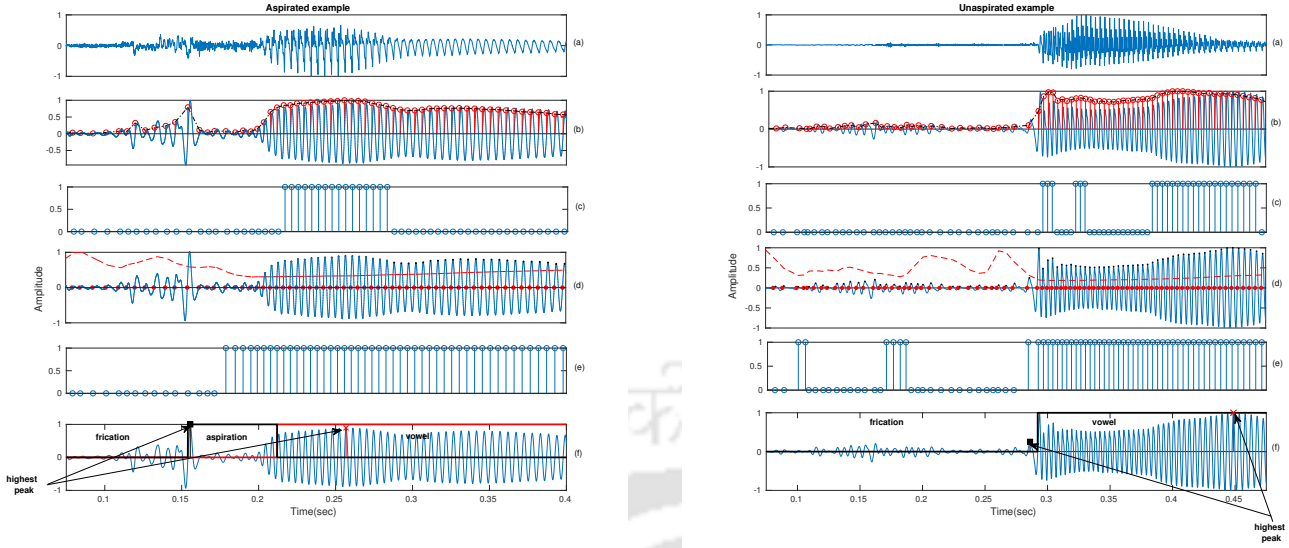

Figure 5.3: Illustration of the proposed method for (i) Aspirated fricative and (ii) Unaspirated fricative (a) Speech signal sampled at 16 kHz (b) Excitation strength plot (dashed black line) of the ZFF signal (c) VLR decision plot with respect to the excitation strength (d) Variance plot (dashed red line) of epoch intervals (red dots); black dots indicate the positive peaks between successive epochs (e) VLR decision plot with respect to '(d)' contour (f) Vowel and non-vowel like region; red box indicates vowel region; black square and red cross indicate the maximum peaks in NVLR and VLR, respectively, VLR represents vowel-like region, NVLR represents nonvowel-like region

$$\alpha + \beta + \tau = 1 \tag{5.5}$$

In (5.4) S_h, S_r denote the scores obtained from the hybrid and random forest classifiers, respectively. The value of α, β, τ is varied from 0 to 1 in steps of 0.01, and the performance is calculated using these weight parameters. The score level fusion of the probability estimates set with optimal weights results in the hybrid classifier. Each fold will select a different best model with varying weight parameters (α, β, τ) assigned to each classifier. The average of the best model results of each fold is computed to estimate the hybrid model's predictive performance.

5.3 Results and Discussion

Previous research has explored an approach to detect aspiration in aspirated fricatives in the Rabha language using SoE and VSEI acoustic features. In this chapter, the proposed approach has been explored for the Korean fricatives, and the illustration of the method is shown in Fig 5.3. The SoE and VSEI contours obtained from ZFFS can be seen in Fig 5.3 (b,d), while the decision plot

for the vowel-non vowel region with respect to the features can be seen in Fig 5.3 (c,e), respectively. The voiced regions have a higher strength of excitation values with low variance. Hence, these areas are marked voiced. The voiceless regions are characterized by the relatively low-amplitude values of the SoE and higher variance. As observed, an aspirated fricative is characterized by a large peak in the non-vowel region of the ZFFS, which is not seen in an unaspirated fricative. Hence, the final plot, Fig 5.3 (f), shows three regions– frication, aspiration, and vowel for the aspirated fricatives. In contrast, only two regions, i.e., frication and vowel, are observed in the case of the unaspirated fricatives. Considering the manually labeled database by experts as the ground truth, the binary classification by knowledge-based approach yielded an accuracy of 90% and 78% for the aspirated and unaspirated class of the Korean fricatives, respectively. The SVM and Bayes classifier using the GMM are trained to combine SoE and VSEI features along with 39-dimensional MFCCs. Along with SoE and VSEI features, relative duration has been used as a feature to differentiate between aspirated and unaspirated fricatives. The performance of these acoustic features along-with MFCCs is evaluated on GMM, SVM, RF classifiers. Also, the performance of the integrated features fed to the SVM-GMM-RF hybrid model are compared with the baseline MFCCs, and the results are discussed in the following section.

5.3.1 Performance Analysis using various classifiers

SVMs trained with 39-dimensional MFCCs with appropriate framing and frequency shift parameters are employed for detecting the two classes of fricatives. A linear kernel-induced SVM classifier has been implemented in previous research [43]. In this work, the train-test set for the Rabha database has been experimented with the non-linear kernel, and better results are observed. The results with different combinations of the features for the SVM classifier are presented in Table 5.1.

When the binary SVM classifier is parametrized with the 39-dimensional MFCCs, SoE, and VSEI parameters (M-S-V), the performance is better than the conventional MFCC based approach (M). For the Rabha database, a 1-2 % increase in performance is observed for the M-S-V approach, irrespective of the kernel used. Out of the several clan groups of the Rabha tribe, Rongdani and Maitori are the two major varieties. The results from the previously studied linear kernel-induced

Table 5.1: Performance analysis using SVM classifier using linear (mentioned within bracket) and non-linear kernel. M, S, V , D represents 39-dimensional MFCC, SoE, VSEI and relative duration features respectively (all the accuracy measures are in % and mean and standard deviation of the accuracy for the nested CV are presented).

Features	Rabha	Korean
M	83.78 ± 6.95 (58.06)	82.51 ± 1.13 (57)
S	53.94 ± 2.18 (57.74)	54.67 ± 3.12 (50.88)
V	51.24 ± 1.42 (57)	53.12 ± 1.47 (51.47)
D	79.55 ± 3.44 (54.29)	76.59 ± 2.33 (51.57)
M+S+V	85.48 ± 7.03 (59.94)	82.90 ± 1.12(63.5)
M+D	87.02 ± 6.27 (60.46)	96.74 ± 1.83 (64.76)
M+S+ V+D	**87.11 ± 6.53(66.18)**	**97.13 ± 0.96(76.19)**

Table 5.2: Performance analysis using GMM classifier M, S, V , D represents 39-dimensional MFCC, SoE, VSEI and relative duration features respectively (all the accuracy measures are in % and mean and standard deviation of the accuracy for the nested CV are presented).

Features	Rabha	Korean
M	73.32 ± 3.91	71.78 ± 6.09
S	50.21 ± 1.40	50.53 ± 2.49
V	49.79 ± 1.29	51.12 ± 1.93
D	53.24 ± 7.54	51.35 ± 1.58
M+S+V	75.50 ± 3.05	74.92 ± 5.84
M+D	78.97 ± 1.87	90.16 ± 2.23
M+S+ V+D	**79.45 ± 3.45**	**90.38 ± 1.74**

SVM classifier [43] for the two dialects, i.e., Maitori and Rongdani varieties of the Rabha speakers, have been averaged and presented in the Table 5.1. The SVM classifier trained with the M-S-V features on the Korean speech corpus showed an improvement compared to the 39-dimensional MFCCs approach using the non-linear kernel parameter.

The classification results achieved by applying the GMM-based methodology to the evaluation set are presented in Table 5.2. The conventional MFCC approach performed lower than the M-S-V approach for both Rabha and Korean fricatives. The random forest classification technique results are presented in Table 5.5. The feature-level combination approach outperforms the standalone MFCC features.

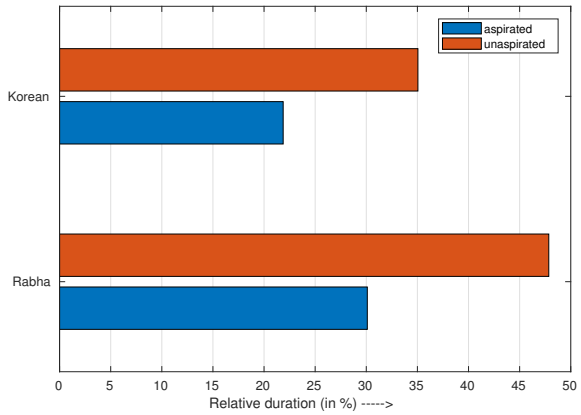

Figure 5.4: Relative duration of the frication segment in the speech samples

5.3.2 Effect of duration feature in the proposed method

As observed in Fig 5.4, the relative duration of the frication segment is longer for the Rabha unaspirated fricative. A similar trend is observed for Korean fricatives. Since the duration evidence distinguishes between aspirated and unaspirated fricatives, the relative duration parameter is taken as another acoustic cue to detect aspiration in the aspirated fricatives. The duration in the aspiration segment is considered as an acoustically distinctive feature associated with Korean aspirated/unaspirated fricatives [41, 61]. The relative duration as an additional feature boosts the classification task's performance by a significant number, consistently for both Rabha and Korean datasets. The reason for such an improvement is discussed later with a t-SNE plot (See Fig 5.5) in the next section.

The M-S-V-D features obtained by appending the relative duration feature to the 41-dimensional feature showed a further increase in the SVM and GMM performance compared to the M-S-V method. For the Rabha database, the binary SVM classifier trained with M-S-V-D features for the two-class classification task showed 87.11 % accuracy when using a non-linear kernel compared to 66.18 % accuracy. For the Rabha language, there is an increase of 4% for the SVM classifier, while an overall increase of 6% is observed for the GMM with respect to the accuracy obtained for the MFCC baseline method. Even for the Korean language, the added features showed an accuracy of 97.13 % and 90.38 % for SVM and Bayes classifier using the GMM, respectively, depicting an overall increase in the performance. The random forest experiment when trained with the M-S-V-D features, too,

showed a similar trend with an accuracy of 83.80 % and 93.61% for the Rabha and Korean datasets, respectively. For all the three classification techniques, there is an overall increase in the M-S-V-D method's performance compared to the conventional MFCC method (as observed in Table 5.5). The developed combined system, i.e., the M-S-V-D method, has exhibited better performance compared to the SVM-based and GMM-based classification using 39-dimensional cepstral features. The better performance of the combined system for differentiating fricatives based on aspiration can be attributed to the capability of acoustic features to capture the additional aspiration information in the fricatives.

The confusion matrices are obtained for each of the train-test sets used in the SVM classifier, and the best performance is shown for the classifiers fed with baseline MFCC features and the proposed M-S-V-D features (See Table 5.3,5.4). With the MFCC features, the Rabha aspirated class is correctly classified with 83.15 % accuracy, while the Rabha unaspirated class is correctly classified with 84.5 % accuracy. The diagonal cells correspond to observations that are correctly classified. The off-diagonal cells correspond to incorrectly classified observations. When the M-S-V-D features are used, the false cases are reduced, thus increasing the overall accuracy of the classifier with 86.99 % correctly classified aspirated cases and 87.4 % correctly classified unaspirated cases. When the binary classifier is tested with the Korean database, the two-class accuracy shows an overall increase in the performance with the M-S-V-D features (96.63 %) compared to the MFCC features (82.19 %). One of the reasons for improving the individual class is that the S-V-D features carry excitation information along with temporal information, which captures some of the other aspiration characteristics that the vocal tract information-carrying MFCC features fails to capture. While producing the unaspirated fricatives, some of the speakers have inadvertently produced some amount of aspiration. Hence those unaspirated examples were misclassified as an aspirated class. In a few examples of aspirated fricatives, the aspiration evidence was not present, and thus they were detected as unaspirated. Error in the manual labeling of the speech segments also contributed to the failure cases.

Table 5.3: Confusion matrices from SVM classifier on the Rabha database with M-S-V-D and MFCC (mentioned within bracket) features

	Aspirated	**Unaspirated**
Aspirated	**86.99% (83.15%)**	13.01% (16.85%)
Unaspirated	12.60% (15.50%)	**87.40% (84.50%)**

Table 5.4: Confusion matrices from SVM classifier on the Korean database with M-S-V-D and MFCC (mentioned within bracket) features

	Aspirated	**Unaspirated**
Aspirated	**96.63% (82.19%)**	3.37% (17.81%)
Unaspirated	3.29% (17.19%)	**96.71% (82.81%)**

5.3.3 Overall study of the features in the two-class classification task for Rabha and Korean fricatives

Based on the experiments conducted, three observations are noticed. First, the non-linear kernel induced SVM and GMM classifier performed better than that using a linear kernel in the previous research [43]. For the Rabha database, the GMM classifier with the M-S-V yields an accuracy of 75.50 %, whereas the 39-dimensional MFCCs yield 73.32 % accuracy. Second, a similar trend is observed in the Korean language, where the conventional MFCC features predict lower accuracy than the M-S-V features. Third, when the relative duration feature is appended to the 41-dimensional feature, the performance of the combined approach is superior to the performance of all the approaches implemented. Also, the overall improvement in the performance is seen irrespective of the classifiers and the data-set used.

Table 5.5 compares the results of the hybrid system with other techniques. The fusion of the SVM, GMM, and random forest classifier at the score level outperforms the individual classifier approach. The results of a confusion matrix for the fusion model constructed with all the three classifiers have been presented in Table 5.6 for the combined M-S-V-D features. The prediction showing the maximum accuracy and the best tuning parameters are chosen for the hybrid classifier. Each classifier is assigned weights, and the hybrid classifier is tuned to the best set of weight parameters (α, β, τ). The accuracy vs. α parameter plot is shown in Fig 5.6. The accuracy for the other two parameters is also investigated similarly (as per Equation **??**, **??**). The optimal weight values are ($\alpha, \beta, \tau = 0.5, 0.2, 0.3$) and $(0.7, 0.1, 0.2)$ for the Rabha and Korean dataset

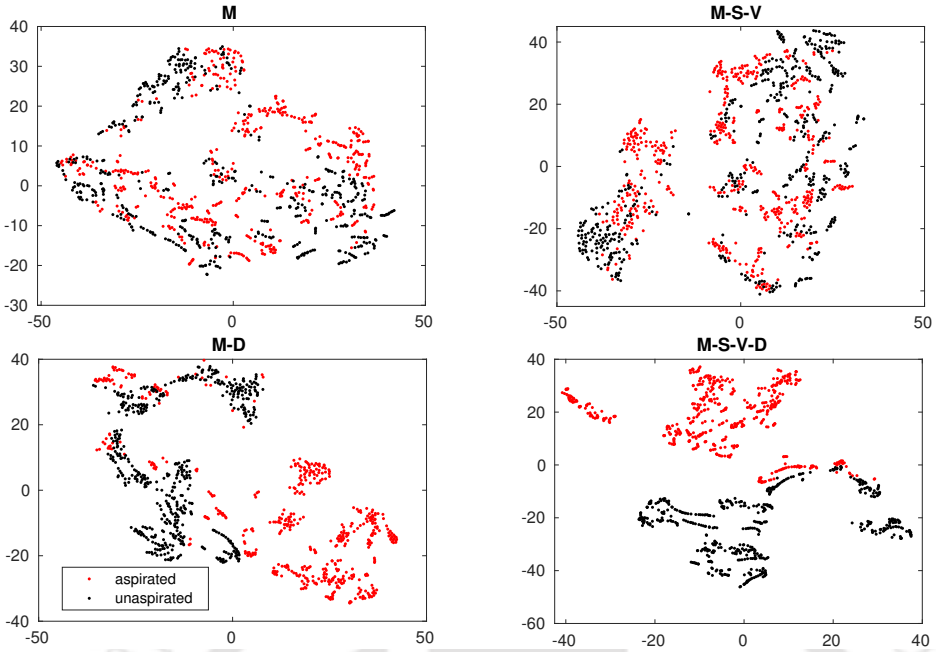

Figure 5.5: Visual distributions of the different features for the Korean database. M, S, V , D represents 39 dimensional MFCC, SoE, VSEI and relative duration features respectively

Table 5.5: Performance analysis using various classifiers M, S, V, D represents 39-dimensional MFCC, SoE, VSEI, and relative duration features respectively (all the accuracy measures are in % and mean and standard deviation of the accuracy are presented).

	four-fold validation				Nested CV			
	Rabha		Korean		Rabha		Korean	
	M	M+S+V+D	M	M+S+V+D	M	M+S+V+D	M	M+S+V+D
GMM	78.06	81.46	76.8	93.61	73.32 ± 3.91	79.45 ± 3.45	71.78 ± 6.09	90.38 ± 1.74
SVM	84.9	89.12	85.31	98.03	83.78 ± 6.95	87.11 ± 6.53	82.51 ± 1.13	97.13 ± 0.96
Random forest (RF)	85.33	89.11	84.74	96.44	73.92 ± 2.53	83.80 ± 3.39	71.90 ± 1.18	93.61 ± 1.61
SVM-GMM	90.61	93.31	99.53	99.69	92.56 ± 5.05	92.92 ± 3.88	98.75 ± 0.72	99.22 ± 0.60
Hybrid (SVM-GMM-RF)	**92.92**	**94.49**	**99.53**	**99.84**	$\mathbf{92.94 \pm 4.95}$	$\mathbf{95.29 \pm 1.90}$	$\mathbf{99.53 \pm 0.31}$	$\mathbf{99.69 \pm 0.36}$

respectively for the hybrid classifier when trained with the M-S-V-D features. The scores of the different classifiers, when fused, capture different attributes of the source, temporal, and vocal tract features, and hence improvement in the performance of the hybrid classifier is observed. Significant improvement in the performance of the combined approach suggests the importance of these features in distinguishing the aspirated- unaspirated class of the fricatives. This result infers that, even though one feature may not provide good performance, when combined, they outperform baseline MFCC irrespective of the classifiers used in the study and efficiently classifies the aspirated-unaspirated variants of fricatives.

The results of the four-fold validation (non-nested) scheme along with the nested cross-validation (CV) strategies on different classifiers are presented in Table 5.5. The generalization error of the

Table 5.6: Confusion matrices of the GMM-SVM-RF hybrid classifier with combined M-S-V-D features

Rabha	Aspirated	Unaspirated
Aspirated	**93.95%**	6.05%
Unaspirated	3.21%	**96.79%**

Korean	Aspirated	Unaspirated
Aspirated	**99.69%**	0.31%
Unaspirated	0.31%	**99.69%**

Table 5.7: Results of the LME model statistical test

	Korean	Rabha
SoE	χ^2 (1, N = 639) =7.27, p <0.01	χ^2 (1, N = 512) =20.26, p <0.001
VSEI	χ^2 (1, N = 639) =43.81, p <0.001	χ^2 (1, N = 512) =36.42, p <0.001
Duration	χ^2 (1, N = 639) = 598.41, p <0.001	χ^2 (1, N = 512) = 29.26, p <0.001

underlying model and its hyper-parameter search is estimated using a nested CV. In the inner CV loop, hyperparameters are tuned using grid-search to get optimized hyper-parameters, while an outer CV is used to classify a held-out test data-set using the optimal value of the hyper-parameters. Choosing the parameters that maximize non-nested CV biases the model to the data-set, leading to over-fitting the training data-set. To overcome optimistically biased evaluation of the model performance, a nested cross-validation strategy is chosen. The goal here is not to compare the nested CV performance with the non-nested one but rather to check how the model trained with M-S-V-D features performs compared to the baseline MFCC approach. It is seen that irrespective of the model's biased or unbiased nature, the hybrid model and all the classifiers trained with the M-S-V-D features outperform the baseline MFCC model.

In order to assess the effect of duration, SoE, and VSEI in fricative aspiration in Korean and Rabha, we conducted an exploratory statistical test by building linear mixed-effects models (LME) for each language. Two separate LME models had duration, SoE, and VSEI as dependent variables and aspiration type (aspirated vs. unaspirated) as a fixed effect. In the case of Korean, speaker, word, and context (phrase vs. isolation) were added in the models as random effects. In the case of Rabha, as the speech data consisted only of isolated tokens, only speaker and word were included in the LME model as random effects. The LME models were built using the *lme4* package on R [143, 144]. In order to see the contrast between the aspirated and unaspirated classes, the LME

models were subjected to a Type II Wald Chi-square test for analysis of deviance using the *car* package on R [145]. The results of the tests are provided in Table 5.7. As seen in Table 5.7, both in Korean and Rabha, aspiration type (aspirated or unaspirated) had a significant effect on all three variables, namely, frication duration, SoE, and VSEI.

The analyses show that accuracy in the classification of aspirated and unaspirated fricatives improves with the addition of the proposed acoustic features, namely, SoE, VSEI, and relative duration (Rdur) to the MFCC features. Hence, we show that the proposed features can be used as parameters to distinguish between aspirated and unaspirated fricatives independent of the language. The combination of source, filter, and temporal features provides better accuracy in the classification of aspirated and unaspirated fricatives.

The visual representation of the different features for the Korean database is shown in Fig 5.5 using t-SNE plots. Fig 5.5 (M) shows a significant overlap between the two categories when only MFCC features are used. Compared to that, adding the SoE, VSEI features to the MFCC features reduces the overlap between the two classes; however, they still do not form distinct clusters. Adding the durational feature to the MFCCs (M-D) reduces overlaps significantly. Finally, when all the four features are combined, as in Fig 5.5 (M-S-V-D), the two categories form more coherent clusters with an increased distance. This indicates that the M-S-V-D features are more effective in distinguishing between the aspirated-unaspirated classes of fricatives. The visualizations here reflect the improved accuracy of the classifiers built with the combined M-S-V-D features compared to the baseline classifier using only MFCC features.

5.4 Conclusion

In this work, MFCC, SoE, VSEI, and relative duration features have been explored as acoustic cues to capture the aspiration information in the aspirated fricatives. The MFCC feature captures the vocal tract information, SoE and VSEI capture the source information, and the relative duration captures the temporal information of the speech sample. While the proposed features should ideally work across languages, it was observed that the Rabha and Korean fricatives are temporally and spectrally quite distinct. Therefore, this work aimed at determining whether or not the proposed features distinguish between aspirated and unaspirated fricatives independent of the language used.

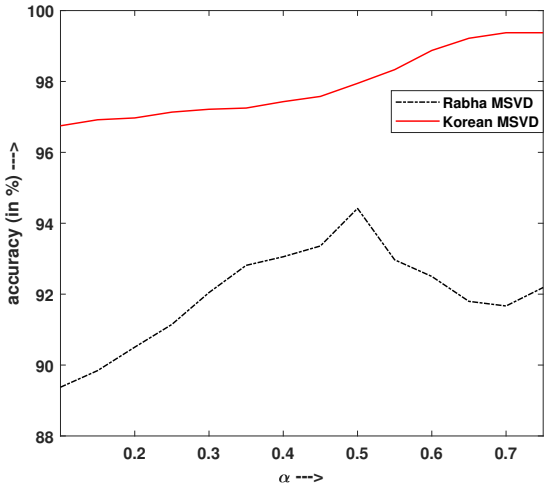

Figure 5.6: Performance curve of the hybrid classifier

This work focuses on the implementation of the proposed features in categorizing the aspirated and unaspirated fricatives using different classifiers. These acoustic features have improved the performance of the system by a significant margin in both Rabha and Korean. The features performed well in the classification task on the two classes of fricatives based on aspiration. Irrespective of the language and the classifier used, the combination of all the acoustic features showed an overall improvement in the performance compared to the conventional cepstral approach.

In the present chapter, aspirated fricatives are distinguished from the unaspirated using fricative specific information. In the next chapter, detection of fricatives and nasals using a common feature is addressed. In particular, next chapter analyzes the spectral dynamics of nasals and fricatives and its effect on classification of two-way contrast non-stop consonants based on aspiration. Irrespective of the consonants used, the feature explored in the next chapter demonstrates whether those features are language independent using multiple classifiers.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

6

Spectral dynamics of aspirated fricatives and aspirated nasals in under-documented languages

Contents

6.1 Introduction

Aspiration is an important phonemic feature that is found commonly in stops and rarely in fricatives and nasals. However, there exists a small body of studies that explored the acoustic characteristics and methods of detection and classification of aspiration in fricatives and nasals occurring in some of the under-documented languages [14, 15, 29, 36]. In a previous work, acoustic features, such as the strength of excitation (SoE), variance of successive epoch intervals (VSEI), normalized autocorrelation peak strength (NAPS) ratio, and vocal tract constriction (VTC) were used to classify aspirated consonants from the unaspirated ones [43]. It was shown in Chapter 5 that using the temporal feature and hybrid classifiers improved classification of aspiration [140]. However, during the production of the aspirated fricatives and nasals, acoustic characteristics were observed to be varying over time. Hence, in the current work, to capture the dynamic spectral changes and to classify the aspirated and unaspirated fricatives and nasals, we employ the equivalent rectangular bandwidth (ERB) based dynamic features [90]. We demonstrate that this feature is able to detect aspiration satisfactorily across different manners of articulation (fricatives vs. nasals) and unrelated languages.

Considering the discussion above, the role of ERB-based features in characterizing the time-varying properties and classifying the fricatives and nasals in terms of aspiration contrast is explored in the current work. Additionally, aspirated and unaspirated classification is attempted on three languages viz. Rabha, Korean, and Angami to see if the ERB-based features are effective cross-linguistically. As the phonemic status of aspirated fricatives in Rabha is not confirmed, a tertiary contribution of this work is a perception study conducted to confirm that Rabha speakers systematically distinguish the aspirated fricatives from the unaspirated ones. The following section provides a detailed discussion on the previous works and motivation and objective of the current work.

6.1.1 Background to the study

From an articulatory point of view, aspiration is considered a puff of air or "turbulent noise" between the articulation of a consonant and the start of voicing, which results in a whispery, breathy, and voiceless speech [6]. When there is an aspiration, the resonances of the vocal tract are excited

by steady airflow from the wide-open glottis [3].

In the case of aspirated fricatives, it is observed that the aspiration portion follows turbulence of the frication portion. It has been reported earlier that frication has a higher energy concentration in the high-frequency regions (≥4000 Hz), whereas aspiration has a higher energy concentration in the low-frequency regions [43]. In the case of voiceless nasals, Angami shows distinctive patterns when there are produced in isolation and in sentence frames. While the ones produced in isolation contain primarily voiceless, nasal, and oral aspiration, in sentence frames, voiceless aspiration is preceded by voiced nasal portions [20]. It was observed that the aspiration region contains components with higher energy (≥2000 Hz or 15 ERB) range as compared to the nasal portion having low energy in that particular region [43]. Hence, both in aspirated fricatives and aspirated nasals, there are dynamic spectral changes over the duration of the consonants. In Fig 6.1(b,f), spectral change is observed between the release of frication and the onset of the following vowel, as compared to the unaspirated variant example of Korean and Rabha fricative. On the other hand, when the oral closure at the lips is released(shortly before 150ms time-point, as seen in Fig 6.1(i)), a period of high oral airflow follows, which overlaps with the onset of the following vowel, i.e., there is "aspiration". The post-release aspiration portion in aspirated nasals has strong nasal airflow, noisy audio signal, that is almost completely voiceless [9]. While the current state-of-the-art is relatively limited when it comes to Angami nasals, aerodynamic studies for the phonetic structures of the voiceless nasals have been conducted for Burmese, Xumi, and Kham Tibetan [15,21]. [20] also noticed that the nasalance of the voiceless nasals drops gradually from about 20% onwards when produced in isolation. On the other hand, when produced in a sentence frame, the nasalance of voiceless nasals reduces sharply from about 60% of the total duration. The noticeable distinction between voiced and voiceless nasals towards the end of the nasals shows that the nasal sounds change over time, and therefore the nasal sound characteristics are dynamic. Hence, to capture the changing dynamics of the fricative and nasal sounds in continuous speech, in the current work, ERB-based features are incorporated.

6.1.2 Motivation and objectives

Several acoustic-phonetic features are proposed in the literature to characterize the aspirated fricatives. Vowel duration, COG, F1 onset, spectral tilt, and vowel intensity are among the few

standard measures to capture the acoustics of aspirated consonants. However, these features may be language-specific. For example, high H1-H2 values are characteristics of Korean aspirated fricatives [52], but the same measure did not show any consistent pattern in characterizing the aspirated fricatives of Bodo and Rabha [16,43]. In the case of Korean fricatives, the first formant values of the following vowels change over time, resulting in significant differences towards the end of the vowels in terms of aspiration [61]. Vowel duration, F1 onset, and vowel intensity characterize the two-way contrast in Korean fricatives, yet it fails to capture the aspiration contrasts in Rabha fricatives and Angami nasals [89]. Among the features explored in the signal processing domain, Vocal tract constriction (VTC) and Normalized autocorrelation peak strength (NAPS) ratio have been capable of classifying aspirated nasals; however, they fail to classify the aspirated fricatives [43]. Features, such as the variance of successive epoch interval (VSEI) and strength of excitation (SoE), are used to distinguish aspiration contrast in fricatives [43]. However, SoE is the only feature that is able to classify both aspirated fricatives and nasals from their unaspirated counterparts. Nevertheless, it is imperative from the discussion above that language-independent, dynamic acoustic features need to be identified to capture aspiration contrasts cross-linguistically.

As discussed above, the fricative and the nasal aspirated consonants do have some acoustic similarity in spite of being articulatorily distinct. Both the consonants are characterized by aperiodic components varying over time. Hence, the motivation is to use a feature that will work for both aspirated fricatives and nasals. In ASR applications, the Gammatone filter bank has been considered superior to the triangular-shaped Mel filters in the presence of noise [91, 146–149]. In previous studies, ERB-based features have been used for noise-like fricative studies. Hence, to better characterize the dynamic noise-like aperiodic changes [91] in aspirated fricatives and nasals, gammatone based equivalent rectangular bandwidth (ERB) feature is used in the current study. In the study of noisy consonants, such as fricatives, it is claimed that the temporal changes in spectral properties are crucial acoustic features [90, 150, 151]. Hence, an ERB-based measure, peak ERB_N, was proposed by [90], which denotes the dominant psycho-acoustic frequency in the production of the fricatives and shows how energy varies according to the frequency [85]. The peak ERB_N is based on a Gammatone filter bank whose center frequencies are distributed according to the ERB scale. Apart from using the ERB-based features, in the current work, multiple machine learning

algorithms were used to compare the performance in classifying aspirated and unaspirated fricatives and nasals in three languages, viz. Rabha, Korean, and Angami.

To summarize, the current work extends the previous work on aspirated fricatives and nasals using an ERB-based feature extraction for studying dynamic spectral characteristics over the time course of the contrastive fricatives, /s/-/s^{h}/ in Rabha and Korean, both phylogenetically unrelated languages. Apart from that, the same feature is also used for characterizing the two-way contrast in Angami nasals /m-m̥/, /n-n̥/ and /ŋ- ŋ̥/ [1]. Also, the knowledge from the differences in the shapes of the two consonants' peak ERB_N trajectories over time is utilized for ERB-based feature extraction and the automatic classification of aspirated and unaspirated variants of fricatives and nasals [152].

The organization of the rest of the chapter is as follows: In Section 6.2, the details of the database collection are briefly stated. A detailed discussion on the importance of analyzing the dynamic spectral characteristics is provided in Section 6.3 and 6.4. The methodology is presented in Section 6.5. Furthermore, the extracted features and the modeling technique for the classification task are described in the experimental setup. In Section 6.6, a discussion of the extracted features and their effect in the binary classification technique of the fricatives in Rabha and Korean is described. A brief overview of the possibility of using the extracted feature for distinguishing aspirated and unaspirated nasals is also presented. Finally, the conclusion and future work are discussed in Section 6.7.

6.2 Database Information

The speech database for acoustic characterization and automatic classification included speech from 44 native Rabha speakers, 19 native Korean speakers, and 21 native Angami speakers. None of the participants reported any history of speech, language, or hearing disorders. The speakers were asked to read the target words [2] in different environments: in isolation and sentence frames.

The speech was recorded with a Tascam DR 100 MKII solid-state audio recorder connected to a Shure SM10 headset microphone. All recordings were conducted in a quiet environment. The speech samples were recorded at a sampling frequency of 44.1 kHz saved in .wav format. Later,

[1] It is to be noted that the phonetic form for the voiceless, bilabial, alveolar and velar nasals are /m̥^{h}, n̥^{h}, and ŋ̥^{h}/. All three voiceless nasals are preceded by a voiced nasal portion when produced in sentence frames.

[2] The target words and gloss have been provided in Table 3.2 - 3.4 in Chapter 3

speech sounds were manually annotated at phoneme level with PRAAT [4] by careful listening and visual inspection of the spectrogram and waveforms. The fricative speech tokens were marked for frication, aspiration, and vowel regions, while the nasals were marked in the phoneme level. For the experiments, phoneme level annotation were used for all the consonants.

6.3 Perceptual evidence for aspiration contrast in Rabha fricatives

While aspiration contrast in Korean fricatives is well-attested [44, 52, 123], such is not the case for Rabha. In order to confirm the phonemic status of aspirated fricatives in Rabha, we conducted a pilot perception study on two subjects, native speakers of Rabha, one male, aged 54, and one female, aged 39 [3]. The perception tests consisted of the /sa/-/s^{h}a/ pair where /sa/ is associated with "to eat" and /s^{h}a/ is associated with "to ache". The stimuli were manipulated using the manipulation tool of PRAAT so that the duration, f0, and intensity of the stimuli were the same. The subjects were asked to participate in a discrimination test where they had to choose whether the two stimuli were the same or different. The test also consisted of control pairs where both the stimuli were the same. The responses of the speakers were measured using the d' score [153]. Following the discrimination test, the subjects were asked to participate in an identification test where the stimuli were presented along with two images representing the meanings of the words. The participants had to identify the stimuli and choose one of the images provided. The participants could discriminate between the /sa/ and /s^{h}a/ with high accuracy ($d' = 2.36$, $d' = 2.69$), confirming their ability to perceive aspiration contrast. In terms of the identification test, the two speakers could correctly associate the stimuli with the meaning with an accuracy of 82.5% and 85%. This confirms that the Rabha fricatives are phonemically distinguished in terms of aspiration.

6.4 Extraction of the features and their statistical tests

This section discusses the systematic experiments for extracting ERB-based features and conventional Mel-frequency cepstral coefficients (MFCCs).

[3]This perception test is conducted on 16 native speakers of Rabha. However, the results from all the participants are not included as it is not the primary objective of this study.

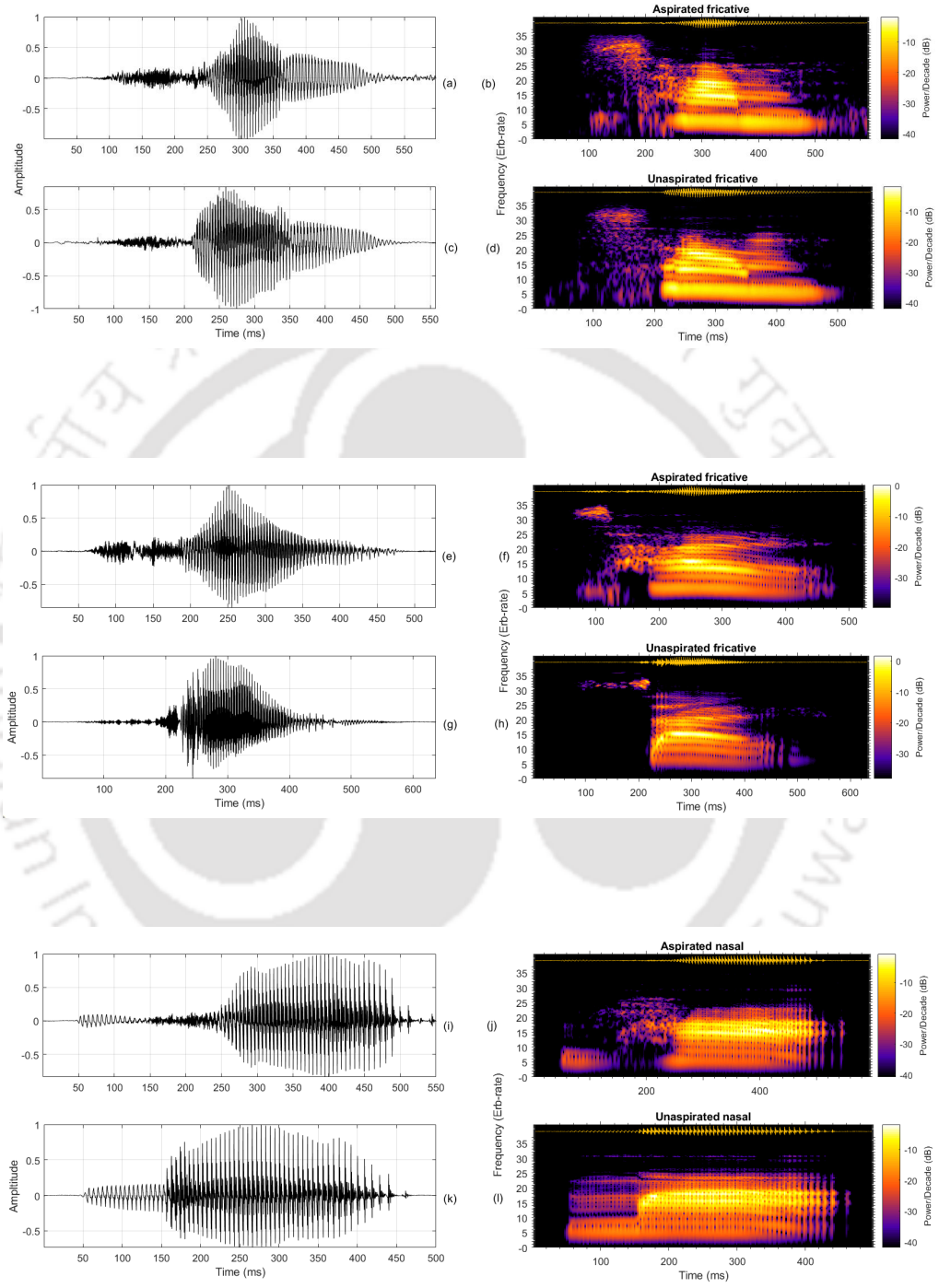

Figure 6.1: Waveform and spectrogram of fricatives (a,b) Rabha /s^{h}am/ (c,d) /sam/; (e,f) Korean /s^{h}aɬ/ (g,h) /saɬ/; nasals (i,j) Angami /m^{h}a/ (k,l) /ma/

In order to obtain the dynamic spectral characteristics of aspirated vs. unaspirated fricatives and nasals, in this work, we explore the ERB-based features as discussed in the earlier section. For the visualization, ERB spectrograms are generated where the frequency values are converted from Hertz to ERBs based on the conversion formula by [154] [4].

In Figure 6.1, the waveforms along with the ERB spectrograms of Rabha (a-d) and Korean (e-h) fricatives are presented. As observed from the figures, frication has a higher energy concentration in the higher frequency regions. In the case of Rabha, the ERB-rate is ≥ 25, whereas, for Korean, it is ≥ 30. The aspiration characteristics of the aspirated fricative are observed in the lower frequency regions for both languages, with an ERB rate of < 25. The Angami nasals are shown in in Figure (6.1 (i-l). The figures show that before the onset of the vowels, the voiceless aspirated nasals have higher energy concentration, with an ERB-rate of ≥ 10. Therefore, for the aspirated fricatives, it is expected that the peak ERB_N values will be lower in the aspiration regions than the frication regions. On the other hand, for the aspirated nasals, the peak ERB_N values will be higher in the aspiration regions than the nasal regions. This aspiration differences observed in the ERB spectrograms serve as the distinguishing feature between aspirated and unaspirated fricatives and nasals. To distinguish the aspirated and unaspirated consonants, ERB-based features are extracted from the potential non-varying regions (P_{nv}), potential varying region (P_v) and entire duration of the consonants (P_t). The frication, aspiration, and nasal regions are estimated based on the nature of peak ERB_N contours.

To study the temporal variation in the spectral characteristics of fricatives and nasals, the manually annotated speech samples were re-sampled to 16 kHz, and the region from the onset of the fricative and nasal (which includes aspiration) to the offset was considered. [5] The entire consonant (fricative and nasal) duration is evenly spaced such that its duration is divided into 100 analysis regions, with a 20 ms duration window with varying frame-shift. The amount of overlap is dependent on the duration of the fricative. Therefore, the frame-shift varies depending on the speech sample duration, with a constant number of analysis regions and window size across all speech samples.

[4] Conversion from Hz to ERB scale is provided in Figure 1 of the Appendix B

[5] Note that in case of aspirated fricatives, the entire length of the phoneme, including both the frication and aspiration regions is considered as the fricative.

To compute the peak ERB-number, the DFT magnitude spectrum calculated from each region is passed through a gammatone filter bank, which transforms the acoustic spectrum into a psycho-acoustic spectrum. The equivalent frequencies in Hz are converted to a row vector of 361 values uniformly spaced on the ERB scale. The filter coefficients contain the coefficients for 361 channel fourth-order gammatone filters defined for simulating the cochlea [85, 91]. Then, the spectral energy at the output of each filter is summed up (termed "auditory excitation"). This auditory excitation represents the spectral envelope for a speech frame of a sound unit. The ERB_N-number corresponding to the maximum summed energy of the auditory excitation is termed peak ERB_N.

The means and standard errors of the observed peak ERB_N of aspirated fricatives /s$^{\text{h}}$/ and unaspirated fricatives /s/ are shown in Figure 6.2 (a,b) for Rabha and Korean, respectively. In the case of Rabha unaspirated fricatives, it is seen that the peak ERB_N values remain in the range of 25-30 throughout the duration of the fricative. However, in the case of the aspirated fricative, the ERB_N values decrease abruptly after 60 % of the total duration.

In the case of Korean fricatives, as seen in Figure 6.2 (b), the peak ERB_N values obtained from the unaspirated fricatives remain in the range of 24-32. On the other hand, for the aspirated fricatives, mean ERB_N values gradually increase to a value of 30 and then decrease abruptly to a value 17 after 40 % of the total duration. This suggests that the aspiration region has a low peak ERB_N frequency in the aspirated fricatives.

In the case of Angami nasals, as observed in Fig 6.3, the peak ERB_N values obtained from the unaspirated nasals remain in the range of 6-10. However, in the case of the aspirated nasals, the ERB_N values increase gradually after 20% of the total duration and abruptly after 40% of the total duration. Hence from all the observations discussed above, a systematic difference is observed between the aspirated and unaspirated fricatives and nasals in terms of peak ERB_N values.

Based on the trends observed in the peak ERB_N contours, features are extracted from the following regions: P_t, P_{nv} and P_v. P_t consists of the 100-dim feature vector computed from the total duration of the speech sample. Second, the feature is extracted from the P_{nv} sound regions. P_{nv} consists of the initial 60-dim feature vector and initial 40-dim feature vector computed from the peak ERB_N number of the fricatives of Rabha and Korean, respectively. P_{nv} are chosen based on the region where mean curve trajectories of the aspirated and unaspirated consonants overlap or

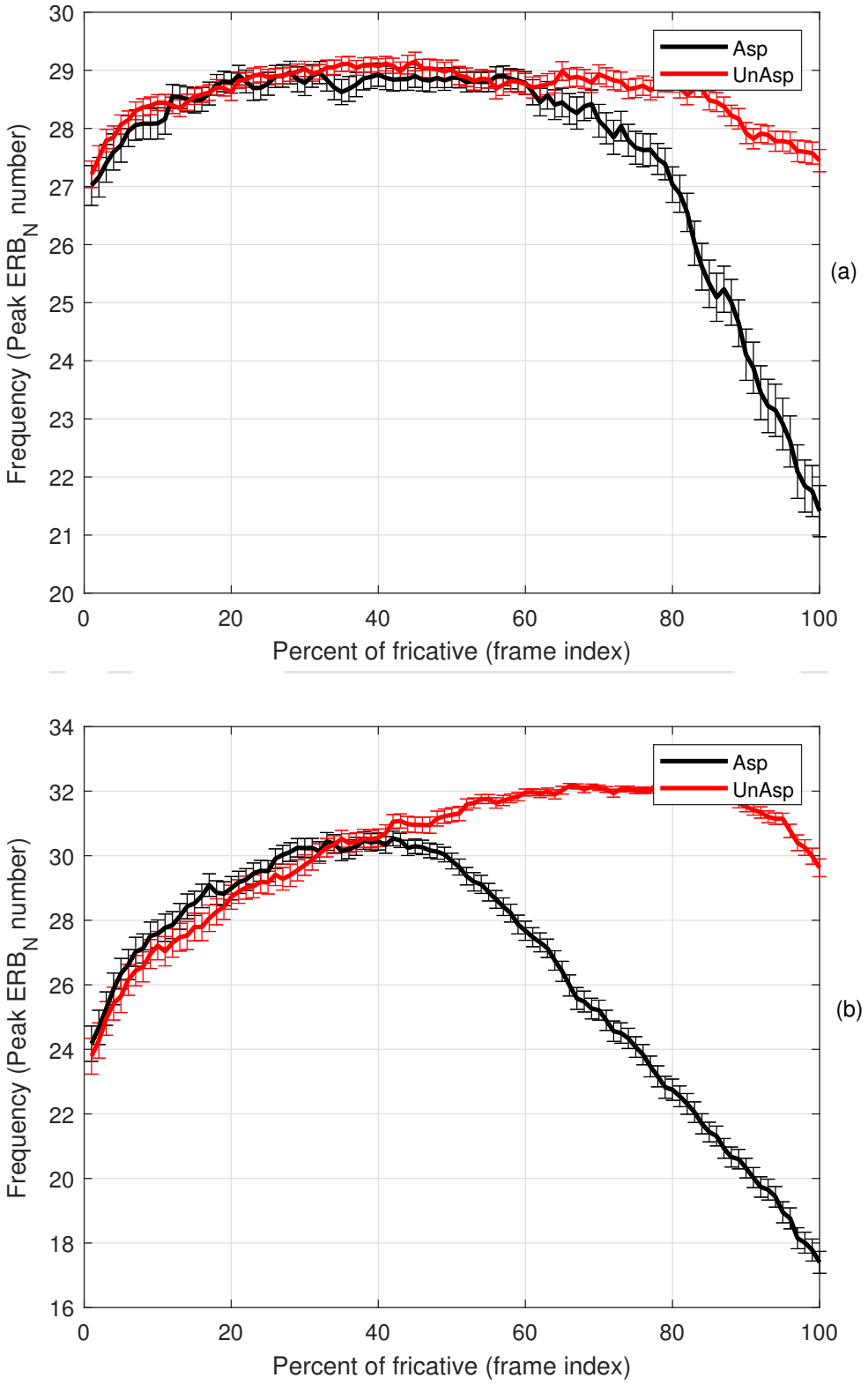

Figure 6.2: Trajectory showing mean and standard error of peak ERB_N values for (a) Rabha and (b) Korean aspirated (black) and unaspirated (red) fricatives.

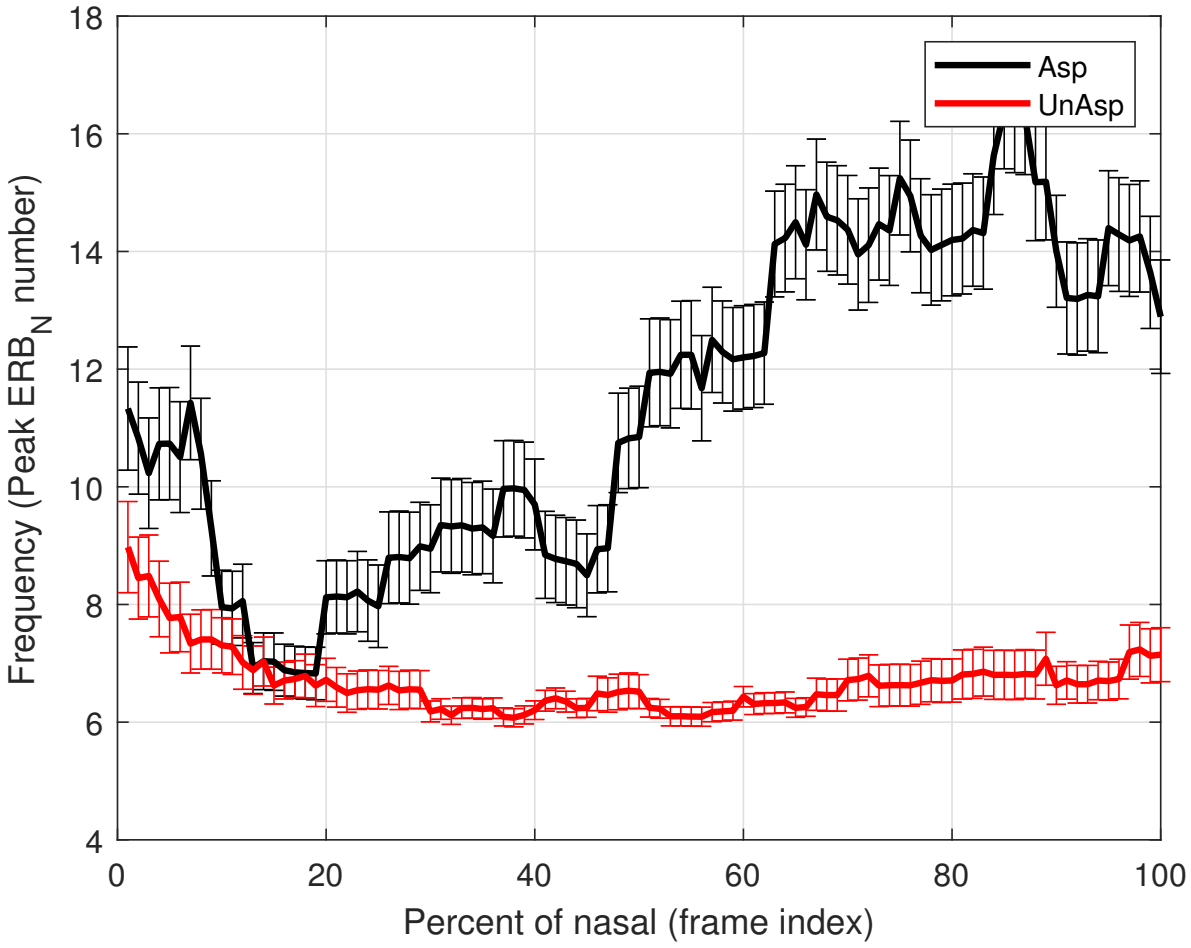

Figure 6.3: Trajectory showing mean and standard error of peak ERB_N values for Angami aspirated (black) and unaspirated (red) nasals.

have the lowest peak frequency difference between the aspirated and unaspirated consonants. P_v consists of the acoustic cues are extracted from the region in which the greatest peak frequency difference is observed between aspirated and unaspirated consonants, i.e., later 40-dimensional feature vector for Rabha fricatives, while later 60-dimensional feature vector is considered for the Korean fricatives. It is hypothesized that the P_{nv} and P_v region carry the frication information and aspiration information, respectively, while the P_t region has the characteristics of the entire time course of the fricatives and nasals.

Apart from the 100-dimensional peak ERB_N extracted as a feature vector, discrete cosine transform (DCTs) coefficients are also computed to model the temporal dynamics. The peak ERB_N contours are subjected to discrete cosine transform (DCTs), and four coefficients, viz., C1-C4, are extracted as a dynamic feature vector. Generally, C1 of DCT models the slope, C2 models the curvature of the trajectory, higher-order coefficients model the high fluctuations. Only 4 (C1-C4) coefficients are heuristically considered to model the ERB_N trajectory in this study. And this 4-dimensional vector is considered as the dynamic feature vector.

6.4.1 Statistics and visualization of the dataset

In this subsection, firstly, results of a statistical analysis is reported that used the multivariate analysis of variance (MANOVA) test. The MANOVA test was conducted to determine the significance of difference between the aspirated and unaspirated consonants in terms of the ERB values obtained at various regions of the fricative and nasal consonants. In the same section, the high dimensional data used in the study representing the two categories of consonants, namely, aspirated and unaspirated, is visualized as two-dimensional t-SNE plots.

To confirm that the ERB-based features of aspirated and unaspirated fricatives are significantly different, a MANOVA test was conducted for aspirated and unaspirated fricatives for both languages. The ERB values at three different parts of speech regions, namely P_{nv}, P_t, P_v, are considered as dependent variables for the statistical test and aspirated and unaspirated are two independent groups. The results of the MANOVA tests are presented in Table 6.2. From the table it is seen that for both Korean and Rabha, the ERB values across P_{nv} region do not differ in a statistically significant way in terms of the aspiration. However, the ERB values for the P_t and P_v portions

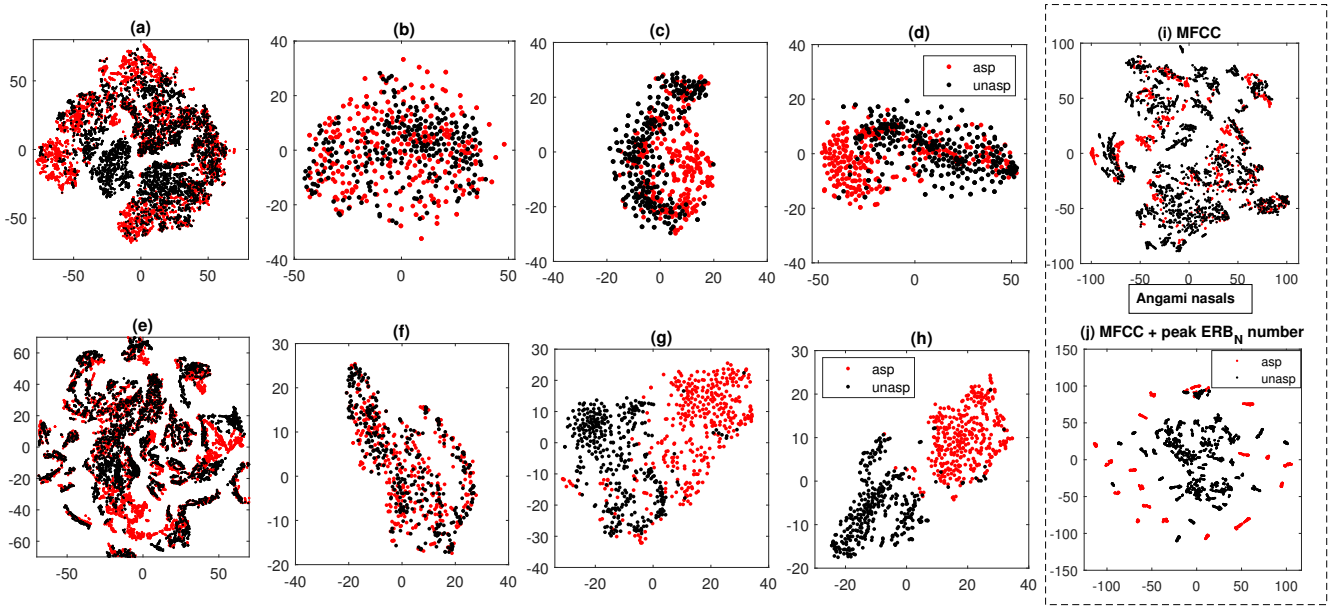

Figure 6.4: T-SNE representation of the cepstral and ERB-based features over P_{nv}, P_t, P_v regions respectively for Rabha (a-d), Korean (e-h) and Angami (i,j).

differ statistically significantly depending on the presence or absence of aspiration.

Unlike the fricatives, a visible difference is observed between the shape of the mean curve trajectories of the two nasals throughout the entire length of the nasals (as seen in Figure 6.3). Statistical analysis showed significant differences for the two groups of nasals when the ERB-based features are extracted from the total length of the nasal speech sample ($p < 0.05$) [6].

A t-Distributed Stochastic Neighbor Embedding (t-SNE) is plotted to visually inspect extracted auditory excitation features, peak ERB_N across the time-course of the fricatives and nasals, as shown in Figure 6.4. The t-SNE is a dimensionality reduction technique that graphically simplifies large datasets with non-linear signals in the data. The t-SNE plot for peak ERB_N values computed from the P_{nv} region indicates that the two-class difference is scattered, and no distinct class difference is observed in Figure 6.4 (b), (f). Therefore, the features extracted from the P_{nv} region may represent some difficulties for the classifiers to differentiate those two sounds based on their acoustic cues. The values for the P_t region visually classifies the two groups to some extent, but many outliers are scattered in each group (in Figure 6.4 (c), (g)). In comparison, the P_v region's peak ERB_N values best discriminate between the two groups with distinct clusters for each fricative class, and

[6]Additionally, statistical analysis of the DCT coefficients extracted have been presented in Table V of the supplementary data in Appendix B but not reported in the current study

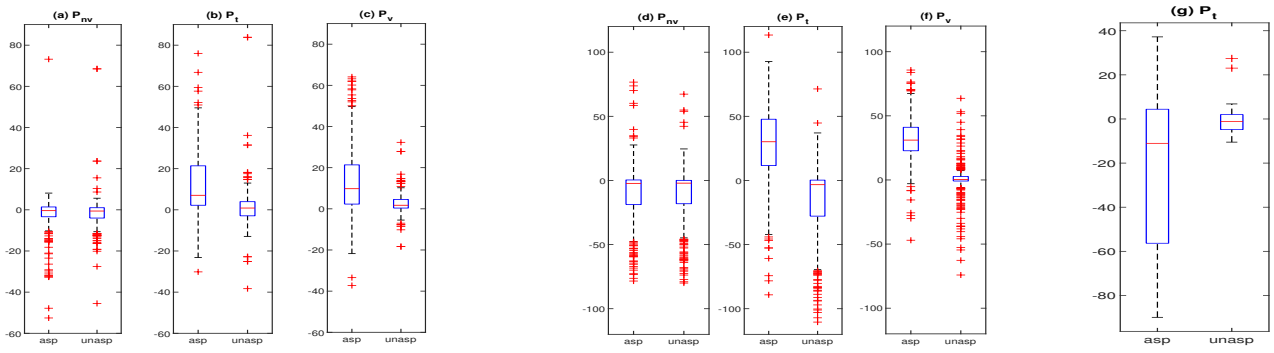

Figure 6.5: Box plot showing the difference in the first coefficient of the DCT (C1) of ERB_N trajectory in terms of aspirated and unaspirated (a-c) Rabha fricatives, (d-f) Korean fricatives and (g) Angami nasals

very few overlap cases are observed in Figure 6.4 (d), (h). In the case of nasals, the dynamic peak ERB-based features, when combined with the MFCCs, show visually distinct clusters compared to only MFCC features (as observed in Fig 6.4(i), (j)).

The MANOVA test and t-SNE plot denote that the P_v region captures the discriminating features between the two fricatives class. In comparison, the P_t region captures the discriminative features between the two nasal classes. Therefore, it is assumed that the ERB_N features with the P_v (fricatives) and P_t (nasals) regions at the front-end of the classifiers will give the best performance. This assumption is verified by evaluating the binary-class classification of the fricatives using an SVM classifier, and the results are presented in Table 6.2 which will be discussed later.

The DCT coefficients are the compressed representation of the peak ERB_N frequency trajectory of the speech frame. Studies have shown the first mean value is consistently lower in unaspirated fricative than in aspirated fricative (also observed in Figure 6.5). The trend observed in the differences of the mean values of the C1 coefficient across the time course of the contrastive fricatives is $P_f < P_t < P_a$. While very little difference is observed in the P_{nv} region that consists of the frication information; for the P_v and P_t region, a noticeable difference is observed in terms of C1 values [7]. In the case of Angami nasals, the first mean value is consistently higher in unaspirated nasals than in aspirated nasals. Hence, the first coefficient of the DCT of the ERB_N values for the fricatives of Rabha and Korean and Angami nasals show that the nature of the ERB_N trajectories of the

[7]In Table IV of the supplementary material in Appendix B, mean and variance of the 1^{st} four DCT coefficients that model the slope and curvature of the peak ERB_N frequency trajectory is presented to study its dynamic nature for Rabha and Korean fricatives.

Table 6.1: Features and mean accuracy computed from the 4-fold classification for Rabha fricatives, Korean fricatives, and Angami nasals. $X^* = P_v$ for fricatives and P_t for nasals respectively.

		Fricatives	Rabha		Korean		Nasals	Angami	
System	**Features**	**Dimension**	**SVM**	**RF**	**SVM**	**RF**	**Dimension**	**SVM**	**RF**
S1	MFCC	39	84.90	85.33	85.31	84.74	39	76.83	74.55
S2	peak ERB_N of X^* region	40 (Rabha)/60 (Korean)	84.64	87.98	96.87	99.16	100	82.36	78.98
S3	DCT coeff computed from peak ERB_N of X^* region	4	78.99	86.17	91.70	95.77	4	82.04	77.78
S4	$S4 = \alpha S1 + (1-\alpha)S_{2,3}$	Score-level fusion	**91.29**	**93.61**	**97.34**	**99.53**	Score-level fusion	**83.09**	**88.36**

aspirated consonants is different from their unaspirated counterparts.

6.5 Acoustic modeling

Two machine learning algorithms, viz. support vector machine (SVM) and random forest (RF) were carried out to evaluate how the extracted features perform in the two-way classification of the fricatives in two different languages. The split for training and testing is 3:1, and they were repeated iteratively and mutually exclusively. An ERB-based feature extraction frontend is integrated into the signal-processing framework of the Rabha and Korean speech corpus. The optimum values of the tuning parameters are experimentally determined using the grid-search method. The grid search algorithm tries all possible combinations of the hyperparameter values and returns the combination with the highest accuracy. This process was repeated for each of the four folds. Average accuracy obtained from the highest accuracy of each of the four folds was used as the overall performance of the machine learning model.

A non-linear RBF kernel is used for the SVM classifier, and the pairs of tuning parameters (C,γ) are evaluated with cost and gamma values being selected within the range 0 to 1, and $exp(-2)$ to $exp(2)$ with a step of 0.1, respectively. The performance of the SVM classifier is tested and compared with the random forest classifier. For the random forest classifier, the two parameters, viz, the number of trees, and maximum features per bag, are tuned within a range of 50-300 and 1-10, respectively, using gridsearch in this study.

The first frontend of the acoustic model consists of the vocal tract features: 12 Mel-frequency cepstral coefficients (MFCCs) and energy with their delta and acceleration coefficients. These 39-dimensional Mel coefficients are represented as M in short in this chapter. The second frontend consisted of the psycho-acoustic gammatone based parameters: peak ERB_N (E) followed by their

Table 6.2: Results of MANOVA conducted on the total duration (P_t) and on regions where the peak ERB_N(E) contours visually differ (P_{nv}) and do not (P_v).

Database	Rabha fricatives	Korean fricatives	Angami nasals
E (P_t)	p<0.001	p<0.001	p<0.05
E (P_{nv})	$p = 0.05$	$p = 0.5$	-
E (P_v)	p<0.001	p<0.001	-

DCT coefficients. For the nasals, too, MFCCs, peak ERB_N feature and their DCT coefficients are extracted from the total length of the speech utterance for the training-testing purpose of the classifiers.

Also, multiple systems are built with different combinations of explored features fed as input to the model. The systems are summarized below: S1 models the MFCCs features, S2 models the peak ERB_N features, S3 models the DCT coefficients, S4 models the weighted score of all the extracted features.

6.6 Results and Discussion

The results of the experiments conducted are presented in this section. In the following subsection, classification accuracy obtained from the two classifiers, SVM and RF, using peak ERB based features are presented. Further, the performance of the various classifiers using these dynamic features are compared with the cepstral coefficients.

6.6.1 Classification using peak ERB based features

The SVM classifier is trained with the peak frequency computed from P_{nv}, P_t, and P_v feature vector to build three models, and the results are presented in Table VI of the supplementary material. It is seen that out of all the models, the ones trained with the ERB values of the P_v region (Rabha: last 40 dimensions, Korean: final 60 dimensions) perform the best in the binary classification task. In the case of Rabha fricatives, an accuracy of 84.64% is observed, while Korean fricatives give an accuracy of 96.87%. The experiments conducted with the RF classifier trained with the features from the P_v region results in an accuracy of 87.98% and 99.16% for Rabha and Korean fricatives, respectively (see Table 7.2, S2). In distinguishing Angami unaspirated and aspirated nasals, accuracy of 82% and 79% were obtained using the ERB-based features for the SVM and RF

classifiers, respectively.

As the ERB_N values of the P_v region perform the best, it is expected that the same would be true for the DCT coefficients too. The experimental results from the three regions showed that the DCT features computed from the peak ERB_N values of the P_v region perform the best amongst all three. Hence, only P_v region's result is reported as S3 system in Table 7.2 . When trained with DCT coefficients, the SVM classifier results in an accuracy of 79% for Rabha fricatives and 91.7% for Korean fricatives. Similarly, the RF classifies with an accuracy of 86.17% and 95.8% for the Rabha and Korean fricatives. The Angami nasals yield an accuracy of 82% and 78% using the DCT features for the SVM and RF classifiers, respectively. This justifies the hypothesis that the discriminating features are captured by the aspiration information-carrying region, as mentioned in Section II.B.

6.6.2 Feature combination system

The classification models are trained with the MFCCs as spectral features. The peak ERB_N values and their DCT coefficients are extracted, which capture the dynamic properties of the peak ERB_N contour. The first four DCT coefficients are used as an additional set of features to train the classification models. Thus, the static and dynamic characteristics are explored to build classification models. The ERB-based features and their DCT coefficients are designed to capture the aspiration distinction in a language-independent manner.

From each of the P_{nv}, P_t and P_v regions, peak ERB_N and their DCT coefficients are extracted to build three separate models for each case. As discussed in the previous section, the model trained with the P_v feature vector performs the best amongst the three. Hereafter, the results computed from the P_v region are reported for all cases of fricatives [8].

Table 7.2 shows the results of the SVM and RF classifiers trained with the acoustic features extracted for the study [9]. The t-SNE plots visualize the relationships between the acoustic cues, and the most visually distinct clusters /s/ and /s^{h}/ sounds are observed for the ERB-based features, as

[8]Table VI of the supplementary material in Appendix B compares the results of all three possible regions and presents the performance evaluation of the psycho-acoustic features for the binary classification task of the fricatives and nasals using SVM.

[9]As significant differences were observed for the two groups of nasals when the ERB-based features are extracted from the total length of the nasal speech sample. Hence the experimental results obtained from the entire duration of the nasals have been presented in the chapter.

seen in Figure 6.4. This can be a possible explanation for the better performance of the ERB-based features compared to that of MFCCs as reported in Table 7.2. The individual scores obtained from the system trained with MFCC (S1), DCT coefficients extracted from the peak ERB_N frequency (S2), and the peak ERB_N frequency features (S3) are combined at score level using z-score normalization technique and weighted score fusion approaches. The weighted measure α value is varied from 0 to 1 in steps of 0.01, and the performance with the optimal weight parameter is reported. This fusion results in the S4 system, and it is seen that the S4 system outperforms baseline MFCCs irrespective of the classifiers or language used (as mentioned in Table 7.2).

For Rabha, the accuracy obtained for the MFCC system is 84.9% and 85.33% for the SVM and RF classifier, respectively. These MFCC results are outperformed by the S4 method with an accuracy of 91.29% and 93.61% using SVM and RF methodology, respectively. For the Korean fricatives, integrating MFCC and all other features for the binary classification task using score-level fusion gives the best accuracy of 97.34% (SVM) and 99.53% (RF) out of all the systems explored.

When trained with the MFCCs, the SVM classifier results in an accuracy of 76.83%, while the RF algorithm classifies with an accuracy of 74.55% for the Angami nasals. The classifier performs better than the conventional MFCCs when trained either with the peak ERB_N features or the DCT coefficients extracted from the total duration of the nasals. The experimental results showed that combining the weighted scores of all the acoustic measures resulted in the best improvement accuracy of 83.09% and 88.36% for the SVM and RF classification approaches.

MFCCs are expected to capture the acoustic properties of a sound unit. However, in the case of acoustically similar two-way contrast consonants, features designed to capture the aspiration distinction perform better than MFCCs [33]. MFCC is sensitive to noise due to its dependence on the spectral form, and hence, it may not be able to discriminate minute differences among the noise-like fricative consonants. On the other hand, ERB is robust to noise, which is one of the reasons for the better performance that was demonstrated in the current work. The components extracted in this work capture the static characteristics and the dynamic characteristics of the two-way contrast fricatives and nasals. Therefore, higher classification accuracy is observed when the system is trained with all the extracted parameters.

6.7 Conclusions

This study investigates the role of spectral features and their dynamic properties in distinguishing aspiration contrasts in fricatives and nasals of Korean, Rabha, and Angami. However, the existence of systematic use of aspiration in contrasting Rabha fricatives was hitherto unconfirmed. Hence, in the current work, through a perception experiment, it was shown that the Rabha fricatives contrast systematically based on aspiration.

Fricative consonants contain significant portions of aperiodic regions. The presence of aspiration changes the aperiodic nature of the fricatives over the total duration of the consonant. In the case of voiceless aspirated nasals, the transition from nasals to aspiration also changes the periodicity characteristics over time. Hence, in this work, it was shown that the variation in aperiodicity could be captured by the peak ERB_N features. The ERB_N features can discriminate the aspirated consonants from their unaspirated counterparts. Irrespective of the classifiers used, the experimental investigation has shown that the combination of the MFCCs and ERB-based features performed the best in discriminating between aspirated and unaspirated fricatives and nasals. In this work, it is seen that the proposed approach of automatic classification of fricatives and nasals contrasting in aspiration performs the best by taking into consideration the region where ERB_N contours differ. However, it is also shown that even when the features are extracted from the entire duration, aspirated and unaspirated consonants can still be well discriminated.

In previous chapters, it was shown that acoustic features, such as SoE, VSEI, VTC, NAPS ratio, do not work uniformly across languages to discriminate between aspirated and unaspirated consonants [43, 140]. In contrast to that, in this work, it was seen that the ERB-based features capture aspiration contrasts effectively in both fricatives and nasals across all three languages. Apart from Korean, these features are also effective in low-resource languages, such as Rabha and Angami.

It was observed that the peak ERB_N trajectories of the aspirated nasals are different from that of the aspirated fricatives. Despite these differences, the ERB_N feature is effective in automatically classifying the aspiration contrast in fricatives and nasals. Hence, this feature may be useful in automatically discriminating non-stop aspiration contrasts in other low-resource languages with a limited amount of speech data.

In the next chapter, different features are combined to demonstrate a framework for the assessment of classification of two-way contrast of fricatives and nasals based on aspiration.

7

Combined framework for the assessment of aspirated non-stop consonants

Contents

7.1 Introduction

In the earlier chapters, the focus on different studies for fricatives and nasals has been made by analyzing the various acoustic-phonetic features. In Chapter 3, a perception experiment was designed to test whether Rabha listeners distinguish and identify the fricatives as aspirated and unaspirated. The discrimination test and identification test of the perception experiment confirm that the Rabha fricatives are phonemically distinguished in terms of aspiration. Further, production analysis on the recorded speech stimuli of the fricatives and nasals was conducted. Results showed that the duration and the COG parameter could be used as a distinctive cue for distinguishing aspirated and unaspirated varieties. Based on the perception and production analysis, acoustic-phonetic information is utilized to extract source features using signal processing methods in Chapter 4. A set of acoustic features is proposed to automatically detect aspiration in fricatives and nasals. Further, in Chapter 5, durational features were used to model a two-class classification system using different classifiers for the fricatives of two languages, Rabha and Korean. Study of varying spectral characteristics over the time course of the contrastive fricatives of Rabha and Korean and contrastive nasals of Angami are discussed in Chapter 6. It was observed that peak ERB_N, effectively classifies the aspiration contrast in fricatives and nasals. However, some missed details in the individual modules may be compensated by combining the individual modules. It is expected that there is a further scope of improvement in the performance evaluation of the aspiration detection algorithm, which is discussed in this chapter.

7.2 Proposed framework

In this chapter, we propose a framework that utilizes all acoustic-phonetic knowledge in a single system to recognize aspirated and unaspirated fricatives and nasals. In this direction, a framework is proposed that can utilize different knowledge at different stages of the framework. The proposed combined framework exploits the independent modules discussed in individual chapters simultaneously, incorporating source, vocal tract, and temporal and spectral information. The proposed framework is shown in Figure 7.1 (a). Also, possible applications of the proposed methods are discussed. In this direction, the effectiveness of the explored features is shown in a phoneme

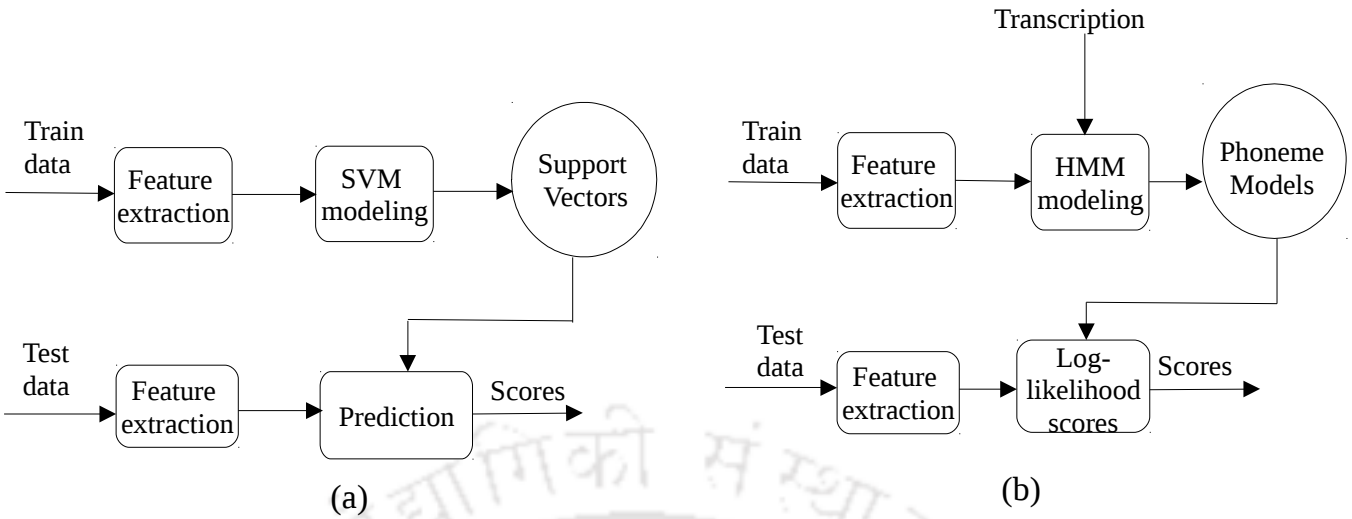

Figure 7.1: Block diagram of (a) SVM classifier (b) Phone recognition system

recognition system which will be discussed later in the subsequent section.

7.2.1 Articulatory motivated feature set in a binary classification task for fricatives and nasals

Initially, a simple classification task is demonstrated using the suprasegmental features that capture the aspiration information in fricatives and nasals. The proposed method uses the two features ($SoE + VSEI$) for Rabha fricatives while three features ($VTC + NAPSratio + SoE$) for the Angami nasals, respectively. Three-fourths of the dataset discussed in Section 5.3 have been used for training purposes and tested on one-fourth of the database. An SVM classifier is chosen for the binary classification task, and the SVM algorithm is implemented in practice using a linear kernel as a baseline method. The developed approach uses three SVM models: first, trained with MFCCs; second, trained with the acoustic features, followed by combined feature set consisting of MFCCs and acoustic features.

The SVM classification results using a non-linear kernel are presented in the next stage. The pairs of tuning parameters (C,γ) are evaluated with cost and gamma values being selected within the range 0 to 1, and $exp(-2)$ to $exp(2)$ with a step of 0.1, respectively. The performance of the SVM classifier is tested and compared with the random forest classifier. For the random forest classifier, the two parameters, viz, the number of trees, and maximum features per bag, are tuned within 50-300 and 1-10, respectively, using grid search in this study. In chapter 6, it was observed that

despite the differences in the articulation of fricatives and nasals, the feature, peak ERB_N, captures the aspiration-specific information in the speech signal and effectively classifies the aspirated non-stop consonants from their unaspirated class. Hence, a framework is developed with this additional articulatory motivated feature along with the previously explored acoustic features (i.e $SoE, VSEI$ for the fricatives; $VTC, NAPSratio, SoE$ for the nasals) by spotting aspiration specific cues to get an optimized acoustic model for the classification of aspirated-unaspirated fricatives and nasals. The knowledge integration framework utilizes the source, vocal tract, prosodic and spectral information and combines all the features to improve the performance of the acoustic model.

Multiple models are trained with different combinations of acoustic features with MFCCs as the baseline model. In the case of the fricatives, S1 models the MFCCs features; S2 models the DCT coefficients computed from the peak ERB_N features; S3 models the MFCCs, SoE, and VSEI; S4 models the relative duration acoustic parameter along-with the spectral cues; S5 model is trained with the weighted score of all the extracted features. In the case of the nasals, S1 models the MFCCs features; S2 models the DCT coefficients computed from the peak ERB_N features; S3 models the three nasal specific acoustic cues, i.e., $VTC, NAPSratio, SoE$; S4 models the MFCCs, $VTC, NAPSratio, SoE$; S5 models the spectral cues; finally, S6 model is trained with the weighted score of all the extracted features.

The proposed framework for each classifier is built using the following equation.

$$S_{proposed} = \alpha S_{baseline} + \beta S_{ss} + \tau S_s \tag{7.1}$$

$$\alpha + \beta + \tau = 1 \tag{7.2}$$

where $S_{proposed}$, $S_{baseline}$, S_{ss}, S_s denote the scores obtained from the proposed framework and baseline features i.e MFCCs, suprasegmental features and segmental features, respectively. The value of α, β, τ is varied from 0 to 1 in steps of 0.01, and the performance is calculated using these weight parameters. The score level fusion of the probability estimates set with optimal weights results in the proposed knowledge integration framework. Each fold will select a different best model with varying weight parameters (α, β, τ) assigned to each classifier. The average of the best model results of each fold is computed to estimate the predictive performance of the proposed

Table 7.1: Performance analysis using SVM (all the measures are in %)

Data	MFCC	Features	MFCC+Features
Rabha (M)	61.08	61.86	63.45
Rabha (R)	55.05	56.8	56.43
Angami	73.18	74.89	76.33

Table 7.2: Choice of features and accuracy computed from the different systems for fricatives. X* = P_v for fricatives.

			Rabha		Korean	
System	**Features**	**Dimension**	**SVM**	**RF**	**SVM**	**RF**
S1	MFCC	39	84.90	85.33	85.31	84.74
S2	DCT coeff computed from peak ERB_N of X^* region	4	78.99	86.17	91.70	95.77
S3	MFCC, SoE, VSEI	41	85.09	87.63	85.93	89.94
S4	relative duration,peak ERB_N , DCT coeff of X^* region	45 (Rabha)/65 (Korean)	87.70	90.09	98.28	99.37
S5	$S5 = \alpha S1 + \beta S3 + \tau S4$	Score-level fusion	94.68	**95.63**	99.69	**99.84**

model.

Table 7.1 shows the two-class classification of aspirated and unaspirated non-stop consonants for Rabha Maitori (M), Rabha Rongdani (R), and Angami database, respectively. Compared to the model trained with the baseline method (39 MFCCs), an overall improvement is observed for all three cases when trained with the acoustic features. When the proposed features are appended with the MFCCs, significant improvement in the performance is observed, which suggests the importance of these features in detecting aspiration in aspirated consonants. In the case of the Rabha Rongdani group, despite an increase in performance using MFCC and the proposed features, the performances for all three cases are consistently near chance level. Analysis of the database suggests that the speakers of this group have a low production of aspiration as compared to the Maitori group speakers, which confirm the literature [110].

For the Rabha corpus, the accuracy obtained for the MFCC system is 84.9% and 85.33% for the SVM and RF classifier, respectively. The proposed S5 model outperforms these MFCC results with an accuracy of 94.68% and 95.63% using SVM and RF methodology, respectively. For the Korean fricatives, integrating MFCC and all other features for the binary classification task using score-level fusion gives the best accuracy of 99.69% (SVM) and 99.84% (RF) out of all the systems explored.

For the Angami database, the system trained with baseline MFCCs reports an accuracy of

Table 7.3: Choice of features and accuracy computed from the different systems for Angami nasals. X* = P_t for nasals.

System	Features	Dimension	SVM	RF
S1	MFCC	39	76.83	74.55
S2	DCT coeff computed from peak ERB_N of X^* region	4	82.04	77.78
S3	NAPS, VTC, SoE	3	73.22	69.20
S4	MFCC, NAPS, VTC, SoE	42	80.89	80.12
S5	peak ERB_N, DCT coeff of X^* region	105	87.04	86.11
S6	$S6 = \alpha S1 + \beta S3 + \tau S5$	Score-level fusion	**87.10**	**86.88**

76.83% and 74.55% for the SVM and RF classifier, respectively. The proposed framework outperforms the baseline model with an accuracy of 87.10% and 86.88% using SVM and RF classifier, respectively.

MFCCs are expected to capture all production and acoustic information of a sound unit. However, the nature of different sound units is different. In the case of acoustically similar two-way contrast consonants, the features designed to capture the aspiration distinction perform better than MFCCs [33]. The features extracted in this work capture the static characteristics and the dynamic characteristics of the two-way contrast fricatives. Therefore, higher classification accuracy is observed when the system is trained with all the extracted parameters.

7.3 Aspiration detection based speech recognition system

Different speech sounds can be characterized by a set of spectral and temporal properties that depend on the acoustic-phonetic features of the sound. When these acoustic-phonetic features are added to conventionally used features such as MFCCs, it enhances the performance of a basic recognition procedure. With the motivation to explore how acoustic-phonetic features perform in a speech recognition system, a phone-level speech recognition system is built for the developed Rabha and Angami speech corpus. The basic block diagram of the proposed phone recognition framework has been illustrated in Figure 7.1 (b).

7.3.1 Angami automatic phone recognition system

In order to see how the aspiration-specific cues contribute to an ASR, we build an ASR with the extracted features and compare it with a baseline method. An ASR is demonstrated for the Angami

Table 7.4: Dataset fed to the automatic speech recognition system

Dataset		Train	Test
Rabha	# utterances	1103	348
	size (min)	87	25

language. Initially, the model is trained with 5709 speech utterances from 7 Angami speakers and tested with 2475 speech utterances from 3 Angami speakers. The train and test set are mutually exclusive. The Angami database used for training-testing the model includes 8184 sentences; each sentence is approximately 2 sec. The recognition system is characterized by utterances resampled to 16 kHz, a window size of 25 msec, and a frameshift of 10 msec. The Angami ASR system is realized using Hidden Markov Model Toolkit (HTK) [155]. A GMM-HMM system is built using HTK and 39-dimensional MFCCs to make the baseline system and compare it with the proposed methodology. GMM-HMM system is built using context-independent mono phone HMMs [2]. To model each of the classes, a 3-state left to right HMM model with a 32 mixture, continuous density diagonal covariance GMMs per state is used.

One set of training models is prepared with a baseline method using 39-dimensional MFCC features. Another set of training models is prepared for aspiration-specific information. In this model, three aspiration detection features are used along with the MFCC features. Conventional 13-dimensional MFCCs, along with their delta and delta-delta coefficients, contain vocal tract information and are helpful for VLRs segmentation as vocal tract shapes are different for VLRs and non-VLRs. Another set of aspiration-specific features is used in the automatic detection method, which includes three aspiration-specific evidence (VTC, NAPS ration, SoE) derived from excitation source and vocal tract information. Thus, a 42-dimensional feature set is used for the aspiration detection-based phone recognition framework. Phone models are built for these sets of features using the HMM-GMM acoustic modeling technique, and performance is evaluated in terms of phone recognition accuracy.

For the Angami speech corpus, an overall accuracy of 73.90 % is observed for the baseline method trained with MFCCs. Significant improvement is achieved for the Angami nasals after adding the 3-dimensional proposed features ($VTC + NAPSratio + SoE$) as seen in Figure 7.2 (a). The proposed modeling technique, which includes the aspiration-specific cues and the MFCCs,

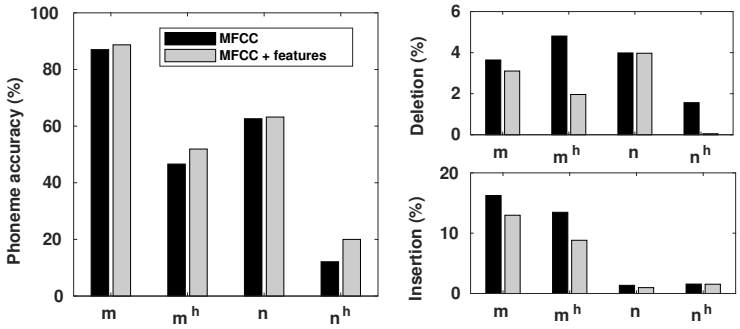

Figure 7.2: (a) Phoneme accuracy before & after appending features for Angami nasals (left) (b) Confusion reduction for Angami nasals (right)

results in an accuracy of 75.58% with an increase of 2% compared to the baseline method trained with MFCCs. The combinations of auditory-based and articulatory-based features have yielded significant improvement in the performance of the recognition system over the baseline feature.

It has been observed that when the model is trained with the baseline method, it yields lower accuracy compared to the proposed method. One of the reasons for the lower accuracy is that aspirated nasals are confused with the phoneme /h/ when trained with MFCC feature-based classification. However, such confusion is reduced when appending MFCC features with aspiration-specific features, i.e., VTC, NAPS ratio, and SoE. Thus, when the model is trained by combining all the features, it reduces both deletion and insertion errors. As such, improved aspirated and unaspirated nasal classification in Angami is observed as seen in Figure 7.2 (b). Therefore, the proposed aspiration-based Angami phoneme recognizer outperforms the baseline method.

7.3.2 Rabha automatic speech recognition system

The Rabha speech corpus developed for training-testing the automatic speech recognition (ASR) system includes 1219 sentences and 232 isolated words; each sentence is approx. 3 sec. The ASR system is realized using Kaldi Toolkit. The speech corpus is characterized by 52 unique sentences and 265 unique words. The speech utterances are resampled to 16 kHz. The training set of the recognition system consists of 1103 speech utterances from 17 speakers, and the evaluation set consists of 348 speech utterances from 5 speakers. The train and test set are mutually exclusive.

The ASR system is realized using Kaldi Toolkit. The default setting of the Kaldi toolkit was used for feature extraction as well as acoustic models. Mel Frequency Cepstral Coefficients (MFCC) were

used as the baseline features in this work. The MFCC features were computed from speech frames of 25 ms duration with frameshift of 10 ms using a Hamming window and pre-emphasis factor of 0.97. For each frame, one set 13-dimensional MFCC feature vector was computed. A 13-dimensional MFCC feature vector represents the smoothed log power spectrum, along the Mel frequency scale, of a single speech frame. Therefore, to capture the dynamic information of the speech signal, the differential (Δ) and the acceleration ($\Delta\Delta$) coefficients were calculated and appended to the previously obtained 13 features. As a result, 39-dimensional feature vectors were used for training acoustic models. Another set of training models is prepared for the custom features capturing aspiration-specific information (i.e., SoE and VSEI). These features were extracted in Matlab and appended to MFCC features in Kaldi format. The context-independent phones (denoted by 'Mono') were modeled by hidden Markov models (HMM) with three emitting states. Separate acoustic model are introduced for each feature: MFCC, SoE, VSEI. One set of the training model is prepared with a baseline method using 13-dimensional MFCC features. During the testing phase, a trained ASR system was fed with test data. The output of the decoder was compared with the reference transcription in order to compute the errors at word level. The baseline recognition system is trained with bi-gram language model and the performance is presented in terms of WER (word-error-rate). The performance of speech recognition system can be evaluated in terms of Word Error Rate (WER), defined as

$$WER(\%) = 100 * (D + S + I)/N \tag{7.3}$$

Where D is the number of word deletions, S is the number of word substitutions; I is the number of words inserted by the decoder, and N is the number of the words in the reference transcription [156].

Table 7.5 summarizes the experimental results of state-of-the-art, auditory-based features, i.e., MFCC, and their combination with aspiration-specific articulatory custom features. Experiments on the mono phone model of the developed Rabha speech corpus show that combining different acoustic features can improve the accuracy of automatic speech recognition systems. As expected, the WER decreases with the increasing sophistication of feature extraction and of acoustic models. The word error rate is reduced from 8.93% to 7.61% when the mono phone model is trained with the combination of extracted features on the Rabha speech corpus. The combinations of auditory-based

Table 7.5: Word error rates obtained by the combination of baseline features (MFCC) and the aspiration specific custom features (SoE, VSEI) on the evaluation set of the developed Rabha and Korean corpus.

Features	MFCC	MFCC + custom features
Rabha	8.93	7.61
Korean	9.33	7.67

and articulatory-based features have yielded significantly better word error rates when compared to systems optimized on a single feature.

7.3.3 Korean automatic speech recognition system

A baseline ASR in Korean was built with the motivation to observe how the aspiration-specific features explored in this thesis perform in the recognition system as compared to the state-of-the-art auditory-based MFCC features. Experiments for acoustic feature combination have been carried out on a small-vocabulary speech recognition task. The Korean speech corpus developed for the recognition task includes 855 isolated words recorded from seven native Korean speakers. The evaluation set consists of a small vocabulary with 100 isolated words. The isolated words are minimal pairs of Korean fricatives contrasting in aspiration, which were mentioned earlier in Chapter 3. The aspirated and the unaspirated fricatives are considered as two separate phones for training the HMMs.

The default setting of the Kaldi toolkit was used for feature extraction as well as acoustic models. Separate acoustic models have been developed using 13- dimensional MFCCs as baseline features and SoE and VSEI as the custom features that capture aspiration-specific information in the phones. SoE and VSEI evidence is processed framewise with a 25 msec window size with a shift of 10 msec, similar to the feature extraction method of MFCCs. These features are concatenated to the MFCCs in the Kaldi format. Further, different models are trained with various sets of features, and the performance of the phone recognition system is evaluated for each model. HMM, models the temporal properties of the speech signal in all cases. The open-source Kaldi toolkit is used for building the phone recognition system.

The performance for different phones in the small-vocabulary Korean dataset, trained with the bigram language model, is evaluated for baseline MFCCs and proposed methods that combine the

MFCCs with the articulatory-based features. The joint feature combination approach was found to be more robust than the MFCC baseline with consistent improvements in WER. The optimized feature combination method outperformed the baseline method, thus, yielding a WER of 7.67% compared to the WER of 9.33% with baseline features (mentioned in Table 7.5).

This speech recognition task enables an analysis of the errors and provides some insight into which features classify the speech sounds correctly. The improvement in the performance of the recognizer is in terms of reduction of substitutions and insertions. The introduction of the new evidence along with the baseline MFCCs helps in the detection of some units which were undetected earlier.

7.4 Summary

In this thesis, several acoustic features and their combination was studied. In this chapter, different features that capture the aspiration information are combined to demonstrate their application in the aspiration detection-based phone recognition system as well as an automatic speech recognition system. An integrated framework has been developed for the assessment of aspirated non-stop consonants. It was shown that articulatory information could improve the accuracy of the state-of-the-art speech recognition system. The potential of the aspiration-information capturing acoustic-phonetic features is demonstrated by developing a binary classification system using different classifiers and compared with state-of-the-art methods. The proposed methodology captures the source, spectro-temporal and vocal tract information and tackles the research problem of the lack of an automatic system for analysis and detection of unique sounds such as aspirated fricatives and aspirated nasals in two under-developed low-resource languages of North-East India. The next chapter summarizes the various contributions of the present thesis.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

8

Summary and Conclusion

Contents

Ψ

8.1 Summary of the Work

The work presented in the thesis aims to investigate a set of acoustic features and classification methods for the classification of fricative consonants differing in terms of aspiration. In order to understand the aspirated fricatives and aspirated nasals, a brief study on concepts of fricatives, nasalization, and aspiration in stop consonants is presented, and how the challenges of automatization of detection of aspirated fricative and aspirated nasals are different from stop consonants is also demonstrated.

8.1.1 Database, production, and perception of aspirated fricatives and aspirated nasals

Initially, a detailed discussion about the development of a speech database for the thesis work and the procedure of database preparation is provided. In order to study the aspirated and unaspirated fricatives and nasals, the database was created for the Rabha and Angami as there is no standard build-in database available in the public domain for those languages. Data was collected from the Korean speakers as no open-source Korean database was available that was relevant to our area of interest. While the reviewed literature presents Korean fricatives as two-way contrast fricatives: aspirated and unaspirated [14, 15, 29, 34, 35, 41, 44, 51, 52, 61–63] and the presence of the two-way contrast of the Angami nasals have been reported [15, 19, 20, 22], no study has been attempted for the phonological fricative contrast in Rabha. As such, the phonemic and phonological status of the Rabha fricatives is not clear. Without any perceptual data in Rabha that verifies that aspiration has any linguistic consequence, the production analysis and, therefore, the automatic classification of the Rabha fricatives would be meaningless. Hence, this perception experiment would give an insight into the existence or non-existence of phonemic pairs of Rabha fricatives, where aspiration may serve as a cue. A perception experiment was designed to test whether Rabha listeners distinguish and identify the fricatives as aspirated and unaspirated. The discrimination test and identification test of the perception experiment confirm that the Rabha fricatives are phonemically distinguished in terms of aspiration. Further, production analysis on the recorded speech stimuli of the fricatives

and nasals was conducted. Results showed that the duration and the COG parameter could be used as a distinctive cue for distinguishing aspirated and unaspirated varieties.

8.1.2 Analysis of excitation source characteristics in aspirated fricatives and aspirated nasal consonants

This work unearths the details of aspiration characteristics from linguistic studies and further utilizes acoustic-phonetic information to extract source features using signal processing methods. A set of acoustic features is proposed to detect aspiration in fricatives and nasals automatically. Acoustic features, such as vocal tract constriction (VTC), normalized autocorrelation peak strength (NAPS), strength of excitation (SoE), and variance of successive epoch intervals (VSEI) are used to detect aspiration in fricatives and nasals. These features are extracted from the zero-frequency filtered signal of the speech sounds, as it preserves the aspiration information. Results show that VTC, NAPS, and SoE can detect aspiration in nasals, whereas SoE and VSEI can detect aspiration in fricatives.

8.1.3 Acoustic phonetic correlates in detecting aspiration in fricatives of Rabha and cross-linguistic comparisons with Korean

This work describes the acoustic-phonetic features associated with aspiration in the fricatives of Rabha and Korean. It is followed by a description of how these features are used to distinguish between aspirated and non-aspirated fricatives in the two languages. The work further implements a machine learning approach to automatically classify the aspirated and unaspirated fricatives, using the acoustic-phonetic features on various classifiers. In this work, speech data was collected from the Rabha and Korean native speakers producing /s/ and /s^h/ phonemes in their respective languages. In order to capture the distinctive aspiration associated with the fricatives, the strength of excitation (SOE), variance of successive epoch intervals (VSEI) along-with durational features were used to model a two-class classification system. We explore the features by combining source, vocal tract, and temporal information to improve the classification of fricatives based on aspiration. The acoustic parameters are fed to support vector machine (SVM), Gaussian mixture model (GMM), and Random Forest (RF) classifiers. Subsequently, we use the antithetic models and combine them into an SVM-GMM-RF hybrid classifier. The integrated features fed to the hybrid model outperform

baseline MFCCs for classifying fricatives in Rabha and Korean, irrespective of the language.

8.1.4 Spectral dynamics of aspirated fricatives and aspirated nasals in under-documented languages

In previous studies, it is observed that the acoustic-phonetic features are usually consonant-specific. This work explores a language-independent feature, peak ERB_N, capturing the aspiration characteristics of aspirated fricatives and aspirated nasals. Fricatives are studied for Rabha and the Korean language, while nasals are studied for Angami. This work investigates the nature of peak ERB_N trajectories to analyze the spectral dynamics across the time course of fricatives and nasals. The differences in the articulation of the two-way contrast non-stop consonants lead to differences in the contours of peak ERB_N in the aspirated and unaspirated variants. Such differences help to propose an efficient approach for analyzing and automatically classifying the aspirated fricatives and aspirated nasals from their unaspirated counterparts. It was observed that the peak ERB_N trajectories of the aspirated nasals are different from that of the aspirated fricatives. The ERB_N feature effectively classifies the aspiration contrast in fricatives and nasals despite these differences. The experimental investigation has shown that the combination of the static and dynamic acoustic parameters improved the classification accuracy compared to the baseline MFCC features.

8.1.5 Combined framework for the assessment of aspirated non-stop consonants

In this work, we propose a framework addressing the issues of utilizing all the acoustic-phonetic knowledge in a single system for recognizing aspirated and unaspirated fricatives and nasals. In this direction, a framework is proposed that can utilize different types of knowledge at various stages of the framework. Also, possible applications of the proposed methods are discussed by demonstrating the details of an aspiration detection-based phone recognition system on the collected corpus.

8.2 Contributions of the thesis

Following are the significant contributions of the thesis towards analysis and automatic detection of aspirated fricatives and nasals.

- Study of the less-studied, rare, and unique sounds: aspirated fricatives and aspirated nasals,

among the world's low-resource languages viz. Rabha and Angami. In particular, this thesis aims to bridge the gap between linguistic and signal processing fields of study by unearthing the details of aspiration characteristics from linguistic studies and further utilizing that acoustic-phonetic information in extracting features using signal processing methods.

- Evaluation and analysis of the two-way contrast fricatives using the extracted features in an under-documented language, Rabha, further extending the work for a well-documented language, Korean, by employing different machine learning techniques. For that purpose, development of the Rabha speech database as well as Korean speech database to capture two-way contrast in fricatives

- This work reports the acoustic characteristics of the aspirated fricatives and aspirated nasals using production and perceptual analysis

- A method is proposed to automatically detect aspiration in aspirated non-stop consonants using source, spectro-temporal, and vocal tract features of speech

- Study of varying spectral characteristics over the time course of the contrastive fricatives of Rabha and Korean and contrastive nasals of Angami. A brief overview on the role of the extracted features in analysis and classification of the two-way contrast Angami nasals (m-m^{h}, n-n^{h}, ŋ-ŋ^{h}).

- A hierarchical framework is developed to automatically detect aspirated fricatives and aspirated nasals in low-resource under-developed languages by combining the proposed features. The importance of a combined framework using the spectro-temporal acoustic measures over conventional MFCCs for automatic detection of aspirated fricatives and aspirated nasals is demonstrated.

8.3 Future Work

Based on the results of this thesis work, this section provides some of the possible future directions for research

- The Rabha and Angami languages have multiple dialects. Hence, from the linguistic point of view, studying the effect of aspiration in various dialect speakers of Rabha and Angami would be interesting. Also, the study can be extended to other languages of North-East India where such phonemic contrasts are observed based on aspiration presence or absence.

- From the signal processing domain, the present work can further be extended into multiple works, some of which are described below. It was observed that the peak ERB_N trajectories for the aspirated fricatives of Rabha and Korean are quite different. This extracted feature not only gives promising results for the automatic classification of the two-way contrast fricatives, but it also successfully distinguishes the Angami aspirated nasals from the unaspirated ones. Therefore, there is some scope in studying the cross-linguistic variation of this measure, and future work is planned in this direction. Since the spectral characteristic varies from frication to aspiration for aspirated fricatives or nasal to aspiration for aspirated nasals, the LSTM model can be explored in the future to learn that temporal variation.

- As the duration feature has been explored in this work and the sequence kernel too utilizes sequence information, it is another dimension worth exploring in the future how this kernel will impact the aspirated-unaspirated classification of fricatives and nasals. While sequence kernel equipped with SVM has shown better performance in speaker verification and speaker recognition, how this kernel would impact the aspirated-unaspirated classification of fricatives and nasals cannot be commented on without further exploration. Further, the impact of background noise on the proposed automatic classification of aspirated/unaspirated is also yet to be explored in this work. It would be interesting to see how the proposed method would behave with background noise and noise-like characteristics in aspiration.

- The work in this thesis has demonstrated that auditory and aspiration-specific features provide complementary information, which results in optimized performance in an Angami phone recognition system, Korean speech recognition system, and Rabha automatic speech recognition system. The work can further be extended to see the performance of these features in a phone recognition system of Rabha language to detect aspirated non-stop consonants in continuous speech automatically. Further, the work can be extended with more data to hybrid

DNN-HMM or Transformer-based ASR. As these Transformer based ASR are data-hungry, exploring the impact of the proposed features in the low-resource language is not feasible at this point; however, it can be explored in the future with more data collection.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

A

Information of the participants of the Rabha database

Table A.1: Information of the participants of Perception experiment

#	Subject	Age	Sex	Education	Languages known	Dialect	Village	Medical Issue
1	R_M1	39	M	12th pass, B.A incomplete	Rabha, Assamese	Rongdani	Rampur	No
2	R_M2	45	M	10	Rabha, Assamese	Rongdani	Rampur	No
3	R_M4	52/55	M	12th pass	Rabha, Assamese,Garo,Bengali,Hindi,English	Rongdani	Tilapara	No
4	R_M5	48	M	6th pass	Rabha, Assamese,Garo,Bengali,Hindi,English	Rongdani	Tilapara	No
5	R_M6	57	M	6th pass	Rabha, Assamese,Garo,Bengali,Hindi,English	Rongdani	Tilapara	No
6	R_M7	73	M	10th pass	Rabha,Assamese,Hindi,Garo,Bengali	Rongdani	Borjuli	No
7	R_M8	61	M	10th pass	Rabha,Assamese,Hindi,English	Rongdani	Darani	No
8	R_M9	43	M	B.Sc pass	Rabha, Assamese, Garo, Hindi, English	Rongdani	Rampur	No
9	R_F1	39	F	12th pass	Rabha, Assamese, Garo, Hindi	Rongdani	Rampur/Chahari	No
10	R_M10	54	M	12th pass	Rabha, Assamese	Rongdani	Rampur	No
11	R_F2	26	F	B.A pass	Rabha, Assamese	Rongdani	Dosorapara Matiya	No
12	R_F3	36	F	12th pass	Rabha, Assamese, Hindi	Rongdani	Rampur	No
13	R_M11	49	M	B.A pass	Rabha, Assamese	Rongdani	Bamundanga	No
14	R_M12	54	M	P.U pass	Rabha, Assamese, Garo, Hindi, English	Rongdani	Barjuli	No
15	R_M13	42	M	B.A pass	Rabha, Assamese, Garo, Hindi, English	Rongdani	Rampur	No
16	R_F5	30	F	12th pass	Rabha, Assamese	Rongdani	Dairong	No

B

Supplementary material for the experiments conducted in Chapter 6

Table B.1: Results of the four DCT coefficients of the peak ERB_N trajectories of aspirated-unaspirated fricatives and aspirated-unaspirated nasals for the total duration (P_t) of the fricatives and nasals. Mean and standard deviation of the DCT coefficients are presented

Database	Rabha			Korean			Angami		
DCT coefficients		Mean	Std. Deviation		Mean	Std. Deviation		Mean	Std. Deviation
C1	asp	12.50278	16.32151	asp	28.32847	29.99187	asp	-25.0245	37.1586
	unasp	0.893994	10.50435	unasp	-16.0425	27.654	unasp	1.361601	16.94062
C2	asp	-12.7025	14.6201	asp	-24.7437	22.58609	asp	3.11637	30.47421
	unasp	-3.70739	6.859594	unasp	-10.8379	17.4965	unasp	3.98203	11.57544
C3	asp	5.504186	11.83379	asp	-1.68681	18.825	asp	8.987515	23.26019
	unasp	-0.24262	5.190319	unasp	-1.7395	14.96899	unasp	1.752854	8.125295
C4	asp	-5.26964	9.317095	asp	-3.72066	14.67623	asp	0.272707	12.66631
	unasp	-1.97869	4.571688	unasp	-4.10767	12.1619	unasp	1.480708	6.006445

Table B.2: F and p value of MANOVA statistical test conducted with first 4 DCT coefficients for aspirated vs. unaspirated groups of fricatives (Rabha, Korean) and nasals (Angami).(P_{nv} refers to the segment having the least difference in the shape of the mean-curve trajectories of the two fricatives/nasals; P_t refers to the full length of the sound file; P_v refers to the segment having the highest difference in the shape of the mean-curve trajectories of the two fricatives/nasals

Database	DCT coefficients of	F	p
Rabha	P_{nv}	2.056	0.085>0.05
	P_t	50.473	0<0.05
	P_v	43.228	0<0.05
Korean	P_{nv}	1.009	0.402>0.05
	P_t	222.838	0<0.05
	P_v	196.140	0<0.05
Angami	P_t	5.395	0.001<0.05

Table B.3: Results of the binary-class classification accuracy obtained using peak ERB_N features (E) and SVM classifier for Rabha and Korean database, respectively (P_{nv}: segment having the least difference in the shape of the mean-curve trajectories of the two fricatives; P_t : entire length of the sound file; P_v: segment having the highest difference in the shape of the mean-curve trajectories of the two fricatives)

Database	Rabha	Korean	Angami
E (P_{nv})	73.49%	66.39%	-
E (P_t)	83.34%	96.86%	82.36%
E (P_v)	84.64%	96.87%	-

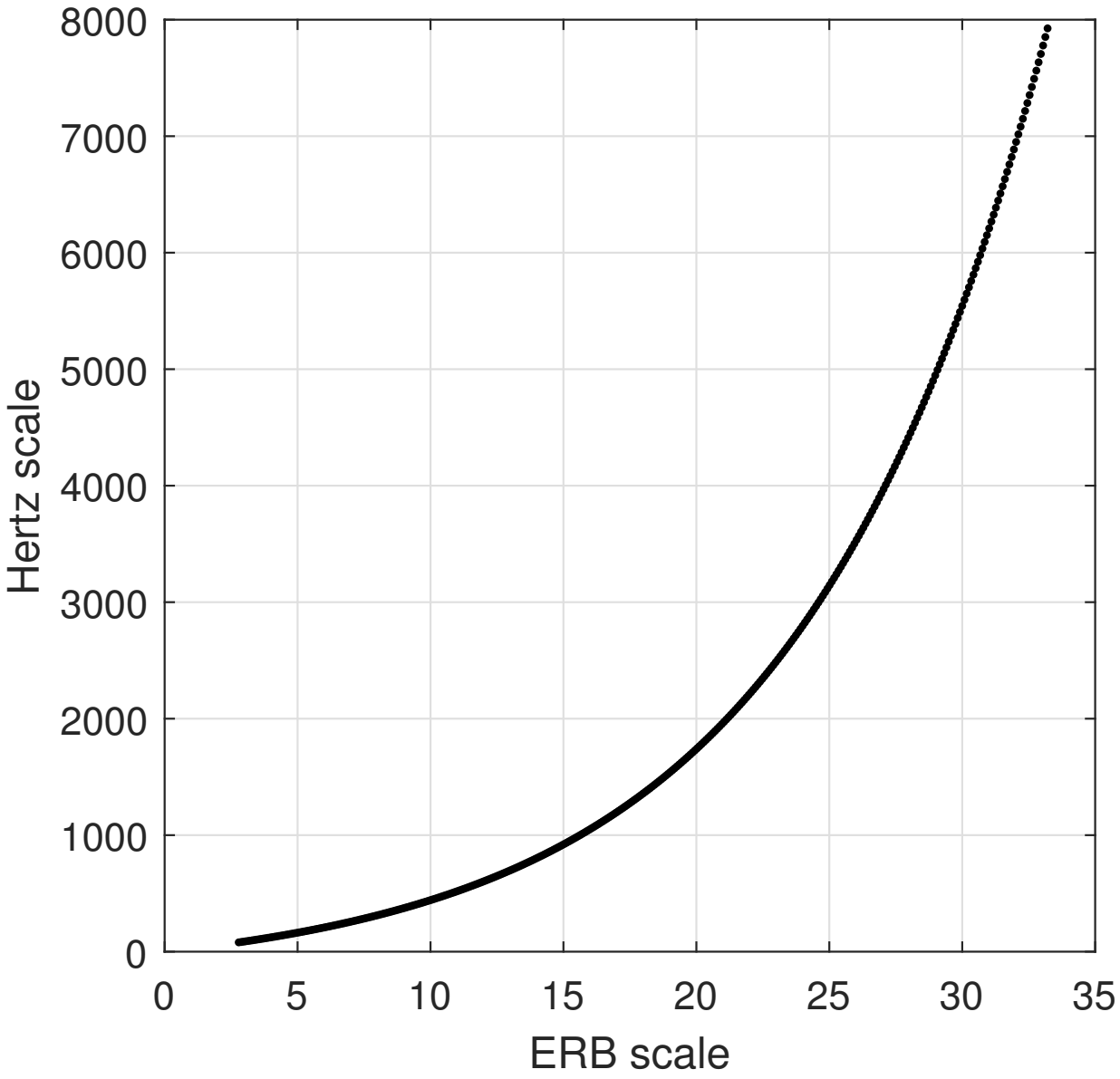

Figure B.1: Conversion of the frequency from Hertz to ERB scale

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Bibliography

[1] J. A. Matisoff, *Handbook of Proto-Tibeto-Burman: system and philosophy of Sino-Tibetan reconstruction*. Univ of California Press, 2003.

[2] B. P. K. Benedict, P. K. Benedict, and J. A. Matisoff, *Sino-Tibetan: a conspectus*. CUP Archive, 1972, no. 2.

[3] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[4] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2001.

[5] B. V. Tucker and R. Wright, "Introduction to the special issue on the phonetics of under-documented languages," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2741–2744, 2020.

[6] J. C. Catford, *A practical introduction to phonetics*. Clarendon Press Oxford, 1988.

[7] J. Schertz and S. Khan, "Acoustic cues in production and perception of the four-way stop laryngeal contrast in hindi and urdu," *Journal of Phonetics*, vol. 81, p. 100979, 2020.

[8] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Blackwell Oxford, 1996, vol. 1012.

[9] I. Maddieson, "Articulatory phonology and sukuma "aspirated nasals"," in *Annual Meeting of the Berkeley Linguistics Society*, vol. 17, no. 2, 1991, pp. 145–154.

[10] P. Ladefoged and D. Everett, "The status of phonetic rarities," *Language*, pp. 794–800, 1996.

[11] D. M. Eberhard, F. S. Gary, and D. F. Charles, "Ethnologue: Languages of the world." *The Journal of the Acoustical Society of America*, vol. Twenty-third edition, 2020. [Online]. Available: http://www.ethnologue.com/

[12] S. Moran, D. McCloy, and R. Wright, Eds., *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2014. [Online]. Available: https://phoible.org/

[13] R. Craioveanu, "The rise and fall of aspirated fricatives," in *Proceedings of the 2013 annual conference of the Canadian Linguistic Association*, 2013.

[14] G. Jacques, "A panchronic study of aspirated fricatives, with new evidence from Pumi," *Lingua*, vol. 121, no. 9, pp. 1518–1538, 2011.

[15] B. Blankenship, P. Ladefoged, P. Bhaskararao, and N. Chase, "Phonetic structures of khonoma angami," *UCLA Working Papers in Phonetics*, vol. 84, pp. 127–142, 1993.

[16] P. Sarmah and P. Mazumdar, "Aspiration in alveolar fricatives in Bodo," in *Proceedings of the 18th International Congress of Phonetic Sciences*, T. S. C. for ICPhS 2015, Ed., vol. 0692. University of Glasgow, 2015, pp. 1–5.

[17] K. H. Berkson, "Acoustic correlates of breathy sonorants in Marathi," *Journal of Phonetics*, vol. 73, pp. 70–90, 2019.

[18] K. Berkson, "Production, perception, and distribution of breathy sonorants in Marathi," *Formal Approaches to South Asian Languages*, 2016.

[19] P. Bhaskararao and P. Ladefoged, "Two types of voiceless nasals," *Journal of the International Phonetic Association*, vol. 21, no. 2, pp. 80–88, 1991.

[20] V. Terhiija and P. Sarmah, "Aspiration in voiceless nasals in angami," in *Proceedings of Meetings on Acoustics 179ASA*, vol. 42, no. 1. Acoustical Society of America, 2020, p. 060008.

[21] K. Chirkova, P. Basset, and A. Amelot, "Voiceless nasal sounds in three Tibeto-Burman languages," *Journal of the International Phonetic Association*, vol. 49, no. 1, pp. 1–32, 2019.

[22] B. Blankenship, "Phonetic structures of an endangered language: Khonoma angami," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2875–2876, 1994.

[23] V. Patil and P. Rao, "Acoustic features for detection of aspirated stops," in *2011 National Conference on Communications (NCC)*. IEEE, 2011, pp. 1–5.

[24] M. Piotrowska, A. Czyżewski, T. Ciszewski, G. Korvel, A. Kurowski, and B. Kostek, "Evaluation of aspiration problems in l2 english pronunciation employing machine learning," *The Journal of the Acoustical Society of America*, vol. 150, no. 1, pp. 120–132, 2021.

[25] M. Shahin and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Communication*, vol. 111, pp. 29–43, 2019.

[26] V. Aubanel and N. Nguyen, "Automatic recognition of regional phonological variation in conversational interaction," *Speech Communication*, vol. 52, no. 6, pp. 577–586, 2010.

[27] Y. Jiao, V. Berisha, J. Liss, S.-C. Hsu, E. Levy, and M. McAuliffe, "Articulation entropy: An unsupervised measure of articulatory precision," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 485–489, 2016.

[28] S. Y. Cheon and V. B. Anderson, "Acoustic and perceptual similarities between English and Korean sibilants: Implications for second language acquisition," *Korean Linguistics*, vol. 14, no. 1, pp. 41–64, 2008.

[29] H. Salgado, J. Slavic, and Y. Zhao, "The production of aspirated fricatives in Sgaw Karen," 2013.

[30] "Online: aspirated consonants," https://en.wikipedia.org/wiki/Aspirated consonant.

[31] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[32] C.-W. Kim, "A theory of aspiration," *Phonetica*, vol. 21, no. 2, pp. 107–116, 1970.

[33] V. V. Patil and P. Rao, "Detection of phonemic aspiration for spoken Hindi pronunciation evaluation," *Journal of Phonetics*, vol. 54, pp. 202–221, 2016.

[34] C. B. Chang, "The production and perception of coronal fricatives in Seoul Korean: The case for a fourth laryngeal category," *Korean Linguistics*, vol. 15, no. 1, pp. 7–49, 2013.

[35] S. Cho and J. Whitman, *Phonology and Phonetics. In Korean: A Linguistic Introduction (pp. 63-95).* Cambridge University Press, 2019.

[36] S. Rabha, P. Mazumdar, P. Sarmah, and S. R. M. Prasanna, "Detection of aspiration in Rabha alveolar fricatives using zero frequency filtering," in *Region 10 Conference, TENCON 2017-2017 IEEE*. IEEE, 2017, pp. 2478–2482.

[37] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[38] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[39] C. X. Xu and Y. Xu, "Effects of consonant aspiration on mandarin tones," *Journal of the International Phonetic Association*, vol. 33, no. 2, pp. 165–181, 2003.

[40] A. S. Abramson and D. H. Whalen, "Voice onset time (vot) at 50: Theoretical and practical issues in measuring voicing distinctions," *Journal of phonetics*, vol. 63, pp. 75–86, 2017.

[41] T. Cho, S.-A. Jun, and P. Ladefoged, "Acoustic and aerodynamic correlates of Korean stops and fricatives," *Journal of phonetics*, vol. 30, no. 2, pp. 193–228, 2002.

[42] G. K. Iverson, "Koreans," *Journal of Phonetics*, vol. 11, no. 2, pp. 191–200, 1983.

[43] S. Rabha, P. Sarmah, and S. R. M. Prasanna, "Aspiration in fricative and nasal consonants: Properties and detection," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 614–625, 2019.

[44] C. B. Chang, "The acoustics of Korean fricatives revisited," *Harvard Studies in Korean Linguistics*, vol. 12, pp. 137–150, 2008.

[45] H. Kim, S. Maeda, and K. Honda, "The laryngeal characterization of Korean fricatives: Stroboscopic cine-mri data," *Journal of Phonetics*, vol. 39, no. 4, pp. 626–641, 2011.

[46] H. Kim, S. Maeda, K. Honda, and S. Hans, "The laryngeal characterization of Korean fricatives: Acoustic and aerodynamic data," *Turbulent sounds: An interdisciplinary guide*, pp. 143–166, 2010.

[47] H. Kim and C.-L. Park, "An acoustic study of the Korean fricatives /s, s'/: Implications for the features spread glottis and tense," *Tones and Features: Phonetic and Phonological Perspectives*, vol. 107, p. 176, 2011.

[48] T.-Y. Jang, "Phonetics of segmental f0 and machine recognition of Korean speech," 2000.

[49] R. Kagaya, "A fiberscopic and acoustic study of the Korean stops, affricates and fricatives," *Journal of phonetics*, vol. 2, no. 2, pp. 161–180, 1974.

[50] H.-S. Park, "The phonetic nature of the phonological contrast between the lenis and fortis fricatives in Korean," *MALSORI*, no. spc1, pp. 24–32, 2002.

[51] K. Yoon, "Study of korean alveolar fricatives: An acoustic analysis synthesis and perception experiment." Mid-America Linguistics Conference, 1999.

[52] T. Cho and S.-A. Jun, "Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean," *UCLA working papers in phonetics*, pp. 57–70, 2000.

[53] K.-S. Kang, "The Korean fricatives in acquisition: A case study," *Speech Sciences*, vol. 11, no. 2, pp. 71–87, 2004.

[54] M. Dantsuji, "Some acoustic observations on the distinction of place of articulation for voiceless nasals in Burmese," *Studia phonologica*, vol. 20, pp. 1–11, 1986.

[55] Dantsuji, Masatake, "A study on voiceless nasals in Burmese." *Studia phonologica*, vol. 18, pp. 1–14, 1984.

[56] W. Lalhminghlui and P. Sarmah, "Characterizing voiced and voiceless nasals in mizo," *Proc. Interspeech 2021*, pp. 426–430, 2021.

[57] V. K. Mittal, B. Yegnanarayana, and P. Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1932–1941, 2014.

[58] R. Lass, *Phonology: An introduction to basic concepts*. Cambridge University Press, 1984.

[59] O. Aziz, "Nasal aspirates in urdu," *Center for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences, Lahore, Pakistan*, 2002.

[60] K. N. Stevens, "Models for the production and acoustics of stop consonants," *Speech communication*, vol. 13, no. 3, pp. 367–375, 1993.

[61] C. B. Chang, "Korean fricatives: Production, perception, and laryngeal typology," *UC Berkeley PhonLab Annual Report*, vol. 3, no. 3, 2007.

[62] J. J. Holliday, "The acoustic realization of the Korean sibilant fricative contrast in Seoul and Daegu," *Phonetics and Speech Sciences*, vol. 4, no. 1, pp. 67–74, 2012.

[63] Holliday, Jeffrey J, "The perception of Seoul Korean fricatives by listeners from five different native dialect and language groups," *Korean Linguistics*, vol. 16, no. 2, pp. 91–108, 2014.

[64] H. Ahn, "An acoustic investigation of the phonation types of post-fricative vowels in Korean," *Harvard Studies in Korean Linguistics*, vol. 8, pp. 57–72, 1999.

[65] V. Patil and P. Rao, "Automatic pronunciation feedback for phonemic aspiration," in *Speech and Language Technology in Education*, 2013.

[66] S. Mikuteit and H. Reetz, "Caught in the act: The timing of aspiration and voicing in east bengali," *Language and speech*, vol. 50, no. 2, pp. 247–277, 2007.

[67] R. Khatiwada, "Some acoustic cues in the detection of the nepalese aspiration," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3889–3889, 2008.

[68] O. Dmitrieva and I. Dutta, "Acoustic correlates of the four-way laryngeal contrast in Marathi stops," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1756–1756, 2018.

[69] Y.-Y. Kong, A. Mullangi, and K. Kokkinakis, "Classification of fricative consonants for speech enhancement in hearing devices," *PloS one*, vol. 9, no. 4, p. e95001, 2014.

[70] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, no. 3, pp. 225–239, 2004.

[71] C. Weinstein, S. S. McCandless, L. Mondshein, and V. Zue, "A system for acoustic-phonetic analysis of continuous speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 54–67, 1975.

[72] O. Fujimura, "Analysis of nasal consonants," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1865–1875, 1962.

[73] Fujimura, Osamu, "Formant-antiformant structure of nasal murmurs," in *Proceedings of the speech communication seminar*, vol. 1, 1962, pp. 1–9.

[74] M. Y. Chen, "Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2443–2453, 1995.

[75] Chen, Marilyn Y, "Acoustic correlates of English and French nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2360–2370, 1997.

[76] P. Mermelstein, "On detecting nasals in continuous speech," *The Journal of the Acoustical Society of America*, vol. 61, no. 2, pp. 581–587, 1977.

[77] D. Ingram, *Phonological disability in children*. Elsevier Publishing Company, 1976, vol. 2.

[78] L. Steels, "Modeling the cultural evolution of language," *Physics of life reviews*, vol. 8, no. 4, pp. 339–356, 2011.

[79] B. D. Sarma and S. M. Prasanna, "Analysis of vocal tract constrictions using zero frequency filtering," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1481–1485, 2014.

[80] P. B. Ramteke, S. Supanekar, and S. G. Koolagudi, "Classification of aspirated and unaspirated sounds in speech using excitation and signal level information," *Computer Speech & Language*, vol. 62, p. 101057, 2020.

[81] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[82] S. Kim, "The interaction between prosodic domain and segmental properties: domain initial strengthening of Korean fricatives and Korean post obstruent tensing rule," *UCLA masters thesis*, 2001.

[83] W.-I. Baik, "On tensity of Korean fricatives (electropalatographic study)," *Speech Sciences*, vol. 4, no. 1, pp. 135–145, 1998.

[84] V. Karjigi and P. Rao, "Classification of place of articulation in unvoiced stops with spectro-temporal surface modeling," *Speech Communication*, vol. 54, no. 10, pp. 1104–1120, 2012.

[85] S. Kalita, P. N. Sudro, S. M. Prasanna, and S. Dandapat, "Nasal air emission in sibilant fricatives of cleft lip and palate speech." in *INTERSPEECH*, 2019, pp. 4544–4548.

[86] S. Kalita, K. Girish, S. Mahadeva Prasanna, and S. Dandapat, "Objective assessment of cleft lip and palate speech intelligibility using articulation and hypernasality measures," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1164–1175, 2019.

[87] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 753–767, 2011.

[88] C. Vikram, N. Adiga, and S. M. Prasanna, "Detection of nasalized voiced stops in cleft palate speech using epoch-synchronous features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1189–1200, 2019.

[89] S. Rabha, P. Sarmah, and S. M. Prasanna, "Acoustic correlates of aspiration in fricatives and nasals," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 986–990.

[90] P. F. Reidy, "The spectral dynamics of voiceless sibilant fricatives in english and japanese," Ph.D. dissertation, The Ohio State University, 2015.

[91] Reidy, Patrick F, "Spectral dynamics of sibilant fricatives are contrastive and language specific," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2518–2529, 2016.

[92] A. M. Abdelatty Ali, J. Van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2217–2235, 2001.

[93] C. Chan and K. Ng, "Separation of fricatives from aspirated plosives by means of temporal spectral variation," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 5, pp. 1130–1137, 1985.

[94] K. Driaunys, V. Rudžionis, and P. Žvinys, "Analysis of vocal phonemes and fricative consonant discrimination based on phonetic acoustics features," *Information Technology and Control*, vol. 34, no. 3, 2005.

[95] A. Frid and Y. Lavner, "Acoustic-phonetic analysis of fricatives for classification using SVM based algorithm," in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2010, pp. 000 751–000 755.

[96] N. Dhananjaya, "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. dissertation, Ph. D. dissertation, IIT Madras, Chennai, India, 2011.

[97] H. K. Vydana and A. K. Vuppala, "Detection of fricatives using s-transform," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3896–3907, 2016.

[98] A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1739–1758, 2008.

[99] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.

[100] A. Juneja and C. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," in *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, vol. 2. IEEE, 2002, pp. 726–730.

[101] A. A. Ali, J. Van der Spiegel, and P. Mueller, "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 2. IEEE, 1998, pp. 961–964.

[102] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.

[103] D. R. Dickson, "An acoustic study of nasality," *Journal of Speech and Hearing Research*, vol. 5, no. 2, pp. 103–111, 1962.

[104] J. R. Glass, "Nasal consonants and nasalized vowels: An acoustic study and recognition experiment," Master's thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and . . . , 1984.

[105] J. Glass and V. Zue, "Detection and recognition of nasal consonants in american english," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11. IEEE, 1986, pp. 2767–2770.

[106] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4733–4736.

[107] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[108] K. N. Stevens, S. E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski, "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," *The Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2979–3000, 1992.

[109] Office of the Registrar General and Census Commissioner, India, "Census of India," https://censusindia.gov.in/2011-Common/Archive.html/, 2011, [Online; accessed 21-Jan-2022].

[110] U. Rabha Hakacham, *Focus on the Rabhas*, Kolkata, India, 2010.

[111] S. K. Chatterji, *Origin and development of the Bengali language*, 1926.

[112] S. Chattarji, *Kirata Jana Kriti*, 1975.

[113] U. V. Joseph and R. Burling, "Tone correspondences among the boro languages," *Linguistics of the Tibeto-Burman Area*, vol. 24, no. 2, pp. 41–55, 2001.

[114] U. Joseph, "Causatives in Rabha," *Linguistics of the Tibeto-Burman Area*, vol. 28, no. 2, pp. 79–106, 2005.

[115] P. C. Basumatari, *An introduction to the Boro language*. Mittal Publications, 2005.

[116] ——, *The Rabha Tribe of North-East India, Bengal and Bangladesh*. Mittal Publications, 2010.

[117] U. V. Joseph, *Rabha*. Brill, 2007.

[118] P. C. Basumatary, "A study in cultural and linguistic affinities of the Boros and the Rabhas of assam," 2004.

[119] U. Rabha Hakacham, *Khurangmuk: A Dictionary of Rabha Assamese English languages*. Kolkata, India: ABILAC, 2015.

[120] U. V. Joseph and R. Burling, *Comparative Phonology of the Boro-Garo Languages*. Central Institute of Indian Languages, 2006, no. 530.

[121] "Korean language," https://www.ethnologue.com, *R*etrieved: 30 August 2019.

[122] J. Shin, C.-y. Sin, J. Kiaer, and J. Cha, *The sounds of Korean*. Cambridge University Press, 2012.

[123] H.-M. Sohn, *The Korean language*. Cambridge University Press, 2001.

[124] P. Keating, "D-prime (signal detection) analysis," http://phonetics.linguistics.ucla.edu/facilities/statistics/dprime.htm, 2004.

[125] E. C. Ucla, C. M. Esposito, S. U. D. Khan, and A. Hurst, "Breathy Nasals and/Nh/clusters in Bengali, Hindi, and Marathi."

[126] V. Carveth *et al.*, "Pathways of development for Qiándōng Hmongic aspirated fricatives," 2013.

[127] C. Harwardt, "Comparing the impact of raised vocal effort on various spectral parameters," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[128] L. Lisker, "Is it vot or a first-formant transition detector?" *The Journal of the Acoustical Society of America*, vol. 57, no. 6, pp. 1547–1551, 1975.

[129] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[130] B. K. Khonglah and S. M. Prasanna, "Speech/music classification using speech-specific features," *Digital Signal Processing*, vol. 48, pp. 71–83, 2016.

[131] V. Mittal, "Analysis of nonverbal speech sounds," Ph.D. dissertation, Ph. D. dissertation, International Institute of Information Technology, 2014.

[132] S. R. M. Prasanna and D. Govind, "Analysis of excitation source information in emotional speech." pp. 781–784, 2010.

[133] B. J. Kim, S.-A. Chang, J. Yang, S.-H. Oh, and L. Xu, "Relative contributions of spectral and temporal cues to Korean phoneme recognition," *PloS one*, vol. 10, no. 7, p. e0131807, 2015.

[134] A. C. Lorena and A. C. De Carvalho, "Evolutionary tuning of SVM parameter values in multiclass problems," *Neurocomputing*, vol. 71, no. 16-18, pp. 3326–3334, 2008.

[135] S. Fine, J. Navratil, and R. A. Gopinath, "Enhancing *gmm* scores using SVM hints"," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[136] J. Liu, Z. Wang, and X. Xiao, "A hybrid SVM/DDBHMM decision fusion modeling for robust continuous digital speech recognition," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 912–920, 2007.

[137] R. Djemili, M. Bedda, and H. Bourouba, "A hybrid GMM/SVM system for text independent speaker identification," *International Journal of Computer and Information Science & Engineering*, vol. 1, no. 1, 2007.

[138] F. Hou and B. Wang, "Text-independent speaker recognition using probabilistic SVM with GMM adjustment," in *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003.* IEEE, 2003, pp. 305–308.

[139] R. Chitturi and J. H. Hansen, "Multi-stream dialect classification using SVM-GMM hybrid classifiers," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU).* IEEE, 2007, pp. 431–436.

[140] S. Rabha, P. Sarmah, S. R. M. Prasanna, and S. Hong, "Aspiration detection in Rabha and Korean aspirated fricatives," *Elsevier Journal on Computer Speech and Language*, in press 2021.

[141] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–8, 2006.

[142] L. Dora, S. Agrawal, R. Panda, and A. Abraham, "Nested cross-validation based adaptive sparse representation algorithm and its application to pathological brain classification," *Expert Systems with Applications*, vol. 114, pp. 313–321, 2018.

[143] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[144] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[145] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[146] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

[147] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–649.

[148] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7204–7208.

[149] V. N. Parinam, C. S. Vootkuri, and S. A. Zahorian, "Comparison of spectral analysis methods for automatic speech recognition." in *INTERSPEECH*, 2013, pp. 3356–3360.

[150] S. J. Behrens and S. E. Blumstein, "Acoustic characteristics of English voiceless fricatives: A descriptive analysis," *Journal of Phonetics*, vol. 16, no. 3, pp. 295–298, 1988.

[151] S. Behrens and S. E. Blumstein, "On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 861–867, 1988.

[152] S. Rabha, S. Kalita, P. Sarmah, S. R. M. Prasanna, and S. Hong, "Spectral dynamics of aspirated fricatives and aspirated nasals in under-documented languages," *The Journal of the Acoustical Society of America (JASA)*, in press 2022.

[153] P. Keating, *UCLA Phonetics Lab*, 2005. [Online]. Available: http://phonetics.linguistics.ucla.edu/facilities/statistics/dprime.html

[154] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[155] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, no. 175, p. 12, 2002.

[156] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Publications Related to Thesis

Journal Publications

1. **Saswati Rabha**, Priyankoo Sarmah and S. R. Mahadeva Prasanna, "Aspiration in fricative and nasal consonants: properties and detection," *The Journal of the Acoustical Society of America (JASA)*, 2019.

2. **Saswati Rabha**, Priyankoo Sarmah, S. R. Mahadeva Prasanna and Seung-ah Hong, "Acoustic phonetic correlates in detecting aspiration in fricatives of Rabha and Korean," *Elsevier Journal on Computer Speech and Language*, (under review)

3. **Saswati Rabha**, Sishir Kalita, S. R. Mahadeva Prasanna, Priyankoo Sarmah and Seung-ah Hong, " Spectral dynamics of aspirated fricatives and aspirated nasals in under-documented languages," *The Journal of the Acoustical Society of America (JASA)*, , (under review).

Conference Publications

1. **Saswati Rabha**, Priyankoo Sarmah and S. R. Mahadeva Prasanna, "Acoustic Correlates of Aspiration in Fricatives and Nasals," in *TENCON*, 2019.

2. **Saswati Rabha**, Priyankoo Sarmah and S. R. Mahadeva Prasanna, "Detection of Aspiration in Rabha Alveolar Fricatives using Zero Frequency Filtering," in *TENCON*, 2017.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati