# Assamese Dialect Identification Using Static and Dynamic Features from Vowel

Hem Chandra Das [1,2,*] and Utpal Bhattacharjee [2]

[1] Department of Computer Science and Technology, Bodoland University, Assam, India
[2] Department of Computer Science and Engineering, Rajiv Gandhi University, Arunachal Pradesh, India
Email: hemchandradas78@gmail.com (H.C.D.); utpal.bhattacharjee@rgu.ac.in (U.B.)
*Corresponding author

***Abstract*—This paper introduces a novel method for identifying Assamese dialects by analyzing the acoustic and prosodic aspects of vowel sounds in speech signals. The distinctive characteristics of these dialects are captured through the use of acoustic parameters such as formants (F1, F2, and F3), as well as prosodic features like energy, fundamental frequency (F0), and duration. To evaluate this approach, a comprehensive vowel speech corpus is collected from native Assamese speakers representing four different dialectal regions. Frame-level statistical features are extracted from vowel sounds, while temporal dynamic features are obtained from steady-state vowel segments. The data collection process involves using a phonetically rich script to record both read and spontaneous speech interactions from speakers of the four dialects. Various classification methods, including three decision tree-based classifiers, i.e., Random Forest (RF), Extreme Random Forest (ERF), and Extreme Gradient Boosting (XGB), are applied to distinguish the four dialects. The performance of each feature, whether static or dynamic, is individually evaluated. The study reveals that the identification of Assamese dialects is influenced by factors such as speech length, intensity, pitch, and formant frequencies. To assess the significance of these features in distinguishing dialects and to measure their combined impact on the identification system, single-factor Analysis of Variance (ANOVA) tests are conducted. Notably, when static features are combined with the Extreme Random Forest (ERF) ensemble model, the overall accuracy of dialect identification reaches 77%. This research demonstrates the efficacy of using acoustic and prosodic features to accurately classify Assamese dialects, shedding light on the subtle variations within them. In summary, this paper provides a robust framework for Assamese dialect identification and contributes to our understanding of dialect discrimination, paving the way for more advanced dialect identification systems.**

***Keywords*—assamese dialect identification, formant frequencies, prosodic features, statistical features, dynamic features**

## I. Introduction

Dialects represent distinct pronunciation patterns within a language, observed among speakers in specific geographic regions. These dialectal variations encompass differences in grammar, phonology, and prosody. Multiple environmental factors, such as socioeconomic class, cultural heritage, geographical location, and educational level, influence the pronunciation variations of speakers [1]. The integration of advanced technologies for classifying and distinguishing dialects has the potential to significantly enhance the performance of interactive speech systems. The presence of dialectal characteristics significantly affects the effectiveness of Automated Speech Recognition (ASR) and Human-Computer Interface (HCI) systems [2]. By incorporating dialect knowledge into pronunciation dictionaries and providing training on acoustic features, the effectiveness of speech-based systems can be greatly improved [3]. Dialect recognition plays a crucial role in various forensic science tasks, including speaker verification, speech verification, and speaker profiling [4]. Integrating dialect recognition into user-machine interactions holds promise for enhancing the overall experience of interaction. Moreover, dialect recognition systems have demonstrated their usefulness in various applications such as dialogue processing, retrieval of spoken documents, translation of spoken language, and efficient conversion of speech into text [5]. Moreover, dialect recognition finds applications in identifying native languages, aiding in medical fields, indexing and retrieving previously spoken materials, supporting the media industry, and more [6].

The Language Identification (LID) issue has garnered considerable interest in the field of speech and language processing, and it can be viewed as a specific case of Automatic Dialect Identification (ADI). Creating a resilient ADI system presents difficulties as it must precisely distinguish dialects that belong to the same language category. Consequently, most existing models are language-specific and struggle to generalize to other languages. This challenge arises from the inherent dissimilarities in pronunciation patterns, phonology, and grammatical structures across different languages [7, 8].

To differentiate between various dialects, researchers have proposed examining the differences in pronunciation of phonemes, consonants, or syllables [9]. Acoustic-phonetic variables with spectral and prosodic properties have proven successful in investigating these differences [10]. This study incorporates acoustic-phonetic and prosodic features for dialect recognition,

Manuscript received June 6, 2023; revised July 21, 2023; accepted September 27, 2023; published February 26, 2024.

comparing the performance of an ensemble method and a single Support Vector Machines (SVM) classifier. Spectral features, such as formant frequencies, are extracted to capture pronunciation variations, while energy, duration, and pitch features represent prosodic distinctions between dialects.

To evaluate the system, a new database is created, comprising speech samples from four distinct dialect groups. Experimental analysis includes the application of different techniques, such as a single SVM classifier, three tree-based classifiers, and an SVM-based ensemble classifier. Cross-dataset evaluation is conducted by considering possible combinations of read and spontaneous speech datasets, enabling comprehensive analysis of the system's performance and results.

Previous studies on Automatic Dialect Identification (ADI) have mainly concentrated on popularly spoken languages, including English, Chinese, Japanese, Dutch, Arabic, Spanish, and various others. However, Indian languages like Hindi, Bengali, Kannada, Punjabi, Telugu, Tamil, Assamese, and others, which can be categorized into Indo-Aryan and Dravidian language families, have received relatively less attention. The research endeavors towards dialect identification in various regional languages spoken in India are still at an early stage of development. This is partially because there aren't many pertinent datasets available for the specific regional languages. There have been only a handful of present-day systems that have tried to analyze and classify dialects by utilizing shorter utterances such as vowels and consonants. Furthermore, there is a lack of studies that have specifically explored the static as well as dynamic behavior of vowels in relation to dialects.

This part highlights some works in dialect processing that consider the overall behavior of vowel utterances. The use of vowel acoustics analysis to identify dialects is yet mostly unexplored. Despite the current knowledge gap, a few noteworthy efforts in this direction show potential for advancing the study of dialect recognition.

The prosodic differences between dialects are expressed by the vowel characteristics energy (loudness), duration, and pitch (F0). Evaluation is performed using read and semi-spontaneous speech samples from four different dialects of Assamese. The study introduces feature extraction techniques such as the mean to capture static phenomena and fluctuations in vowel signals to illustrate the dynamic behavior of vowels. These characteristics are quantified to measure variations in vowel pronunciation. To capture the dynamic behavior, Legendre polynomials of degree five are fitted to each contour, with each coefficient representing formant frequency features within the contour. The same Legendre polynomials are utilized to extract pitch and energy contours, simulating dialectal variations in Assamese. These dynamic contour features exhibit distinctive shapes that capture dialect-specific patterns in vowels.

Significant acoustic relationships are determined through a one-way Analysis of Variance (ANOVA) (Single Factor) test, analyzing the eight Assamese vowels across the four dialects. The classification of dialects based on vowels is performed using three ensemble approaches with multiple classifiers. These ensemble techniques outperform traditional single classifier-based methods. The decision tree method serves as the base classifier, and bagging and boosting approaches are employed to investigate how static and dynamic features contribute to the development of the vowel-based ADI system.

Assam, located in northeastern India, is home to the Assamese language, which is spoken by the inhabitants of the region. Assamese has several regional dialects that differ in pronunciation, syntax, and vocabulary across the state. The four major regional dialects of Assamese are: The section should be organized as:

- Central dialect: Spoken in and around the Nagaon district.
- Eastern dialect: Spoken in the Sibsagar district and its neighboring districts.
- Kamrupia dialect: Spoken in Kamrup, Nalbari, Barpeta, Kokrajhar, and some parts of Bongaigaon district.
- Goalporia dialect: Spoken in Goalpara, Dhuburi, and parts of Bongaigaon district.

Identifying the specific dialect being spoken is essential for developing a universal Assamese voice recognition system that can accurately recognize words spoken in the Assamese language and its various dialects. Understanding the distinctions between these dialects is crucial for effective communication and language processing in the region.

Assamese is indeed the official language of the state of Assam in north-eastern India. It belongs to the Indo-Aryan family of languages and has evolved over time through influences from various non-Aryan languages. The pronunciation of Assamese as "Axamiya" by local speakers is a valid representation of the phonetics in the region. Sanskrit, an ancient language, is considered to be the ancestor of Assamese and many other languages spoken in the Indian subcontinent [11]. Assamese has borrowed vocabulary from Sanskrit and has also incorporated words from other languages it has interacted with during its evolutionary history. Within Assam, there are numerous regional dialects spoken in various regions of the state, along with the overall phonemic diversity of the Assamese language. These dialects can be broadly classified into two primary groups: the dialect spoken in upper Assam and the dialect spoken in lower Assam. The standard colloquial Assamese language predominantly derives from the speech patterns found in upper Assam. Linguist Banikanta Kakati has distinguished Eastern and Western dialects of Assamese based on linguistic similarities. Western Assamese pertains to the language spoken in the area encompassing undivided Kamrup and Goalpara, whereas Eastern Assamese refers to the region extending from Sadiya to Guwahati. It is important to acknowledge that languages and dialects can exhibit variations and sub-varieties, and linguistic categorizations may vary depending on the viewpoints of different researchers or experts.

The Western Assamese dialect displays significant variations between Goalpara and Kamrup speakers, resulting in the emergence of numerous sub-dialects. Linguist G. C. Goswami identified a central dialect situated between upper and lower Assamese. He categorized the regional dialects into three groups: Upper Assamese, Lower Assamese, and Central Assamese. Upper Assamese is spoken from Nagaon in the south to Sonitpur in the north, Lower Assamese is spoken from east Kamrup to Goalpara, and Central Assamese is spoken in Darang in the north, Morigaon in the southeast, and Kamrup in the southwest [12, 13].

However, recent studies by contemporary linguists have established four major Assamese dialect groups: Eastern, Central, Kamrupia, and Goalporia dialects [14]. These studies presented the intra-division depiction of each dialect, as shown in Fig. 1. In the districts of Barpeta, Nalbari, and Kamrup, the Kamrupia dialect is spoken, which includes various sub-dialects such as Barpetiya, Nalbaria, Kamrupia, and South Kamrupia. A notable feature of the Kamrupia dialect is the utilization of stress on the first syllable rather than the second-to-last syllable, as observed in Eastern dialects. This change in stress placement significantly affects the pronunciation of words. As an illustration, in the Kamrupia dialect, the Assamese word for vegetable “gourd” is pronounced as /kumra/, whereas in standard Assamese, it is pronounced as /komora/. Unlike Eastern Assamese, which predominantly uses medial vowels, Kamrupia dialect incorporates additional high vowels.

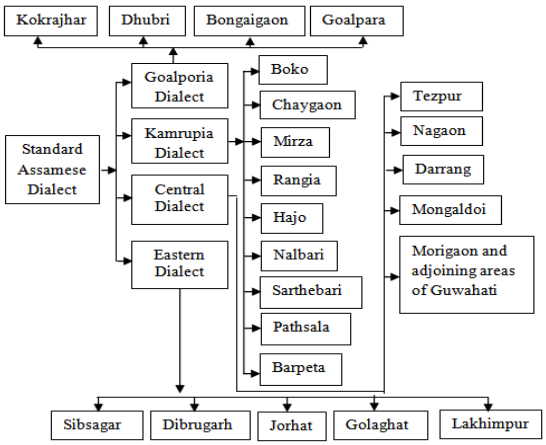


Fig. 1. Assamese dialect’s internal division.

The undivided Goalpara district in Assam, which includes the current Dhubri, Goalpara, Kokrajhar, and Bongaigaon districts, is linked to the Goalporia dialect. This dialect also encompasses eastern and western sub-dialects spoken in the Goalpara, Bongaigaon, and Dhubri regions. Furthermore, the Goalporia dialect shares several morphological and phonological traits with Bengali, another widely spoken language in India [15].

In summary, the Assamese language exhibits variations in dialects, and the categorization of these dialects has evolved over time. The Western Assamese dialects have sub-dialects, and recent studies have identified four major dialect groups: Eastern, Central, Kamrupia, and Goalporia. These dialects differ in stress placement, pronunciation, vowel usage, and linguistic traits, reflecting the rich linguistic diversity within the Assamese language. One of the primary phonological differences among dialects in Assamese often relates to differences in vowels sounds. For example, the word “king” is pronounced as /rɔza/ (ৰজা) in standard Assamese but as /raza/(ৰাজা) in the Kamrupia and Goalporia dialects [16]. These differences in vowel pronunciation contribute to the overall variation observed in the Assamese language across different locations and among different speakers.

The Assamese language is characterized by eight vowel phonemes. Every language possesses a distinct set of vowels that differentiate words from each other. According to scholars such as G. C. Goswami, J. Tamuli, and others, Assamese consists of the following eight vowels: /i/, /e/, /ɛ/, /a/, /u/, /ʊ/, /o/, and /ɔ/ [11, 17]. Among these, the central vowel is represented by /a/, while /i/, /e/, and /ɛ/ are categorized as front vowels, and /u/, /ʊ/, /o/, and /ɔ/ are classified as back vowels. Table I illustrates the organizational structure of the eight Assamese vowels.

TABLE I. EXPLANATION OF VOWEL ORGANIZATION IN ASSAMESE LANGUAGE

| | Front | Central | Back |
|---|---|---|---|
| **High** | i | | u |
| **High** | | | ʊ |
| **High-Mid** | e | | o |
| **Low-Mid** | ɛ | | ɔ |
| **Low** | | a | |

In order to analyse how dialectal differences affect eight monophthongal vowels, the paper seeks to summarize the acoustic-phonetic properties of vowels. By manually segmenting vowels from phonetic-units collected from recordings of both read and semi-spontaneous speech by native speakers, a distinctive dataset of Assamese vowel dialect is created. In this research, an Automatic Dialect Identification (ADI) method based on vowels is presented. In order to do this, the system extracts from vowels energy or loudness or duration information as well as static and dynamic behaviours of spectral formant frequencies (F1–F3) and prosodic parameters like F0 or pitch. The research proposes a feature extraction approach that takes into account features that capture fluctuations in the vowel signals, representing their dynamic behaviour, as well as features that compute mean values to describe the overall static qualities of vowels. To quantify the differences in vowel pronunciation, these characteristics are studied in vowels. These changes can be a result of dialectal differences. The paper uses a Legendre polynomial function of degree five for each contour to capture dynamic behaviours. Within the contour, each polynomial coefficient correlates to particular formant frequency properties. Applying the same Legendre polynomials to represent the dialectal variances in Assamese yields pitch and energy contours as well. These

dynamic contour features have a distinctive form that captures vowel patterns unique to a particular dialect. The fundamental premise of the paper is that vowel qualities in four Assamese dialects exhibit significant acoustic variance. One-way ANOVA (Single Factor) test is used to assess and identify significant acoustic correlations between the eight Assamese vowels from the four dialects. Three ensemble approaches that make use of various classifiers are used to categorise dialects based on vowels. These techniques have proven to perform better than traditional single classifier-based approaches. For the purpose of introducing an ADI system based on vowels, the study focuses on collecting static as well as dynamic patterns in spectral formant frequencies.

## II. Literature Review

This study examines the existing literature on automatic dialect processing, focusing on language models, acoustic-phonetic procedures, phonotactic approaches, and classification methods. The majority of the existing systems for Automatic Dialect Identification (ADI) integrate acoustic-phonetic with phonotactic methods. Scholars have proposed the extraction of dialectal cues from segmental acoustic features (individual speech sounds) and supra-segmental acoustic features (prosodic characteristics). Text-dependent and text-independent speech samples, encompassing both scripted and spontaneous speech modes, are assessed for the purpose for dialect recognition [18]. Researchers have extensively used Mel-Frequency Cepstral Coefficients (MFCC) and Shifted Delta Coefficients (SDC) to analyze spectral acoustic variations between dialects. Gaussian Mixture Models (GMM) are frequently used to capture and analyze spectral cues and temporal variations in order to categorize dialects [8, 10, 19–21]. Moreover, prosodic variations are modeled by extracting rhythmic, intonation, and stress components from pitch, intensity, and duration measurements [22−24]. Certain studies have examined i-vector models that employ joint factor analysis to reduce dimensionality when extracting acoustic features. The inclusion of an i-vector alongside MFCC-SDC features has demonstrated notable enhancements in the performance of dialect recognition [25–27]. The most effective GMM mixtures specific to dialects are generated through the utilization of Kullback-Leibler Divergence-GMM (KLD-GMM) techniques. Moreover, to enhance classification accuracy, Frame Selection Decoding (FSD) is employed by eliminating confounding auditory zones [28].

Several researchers have noted that language boundaries are distinct, and they have suggested the same for dialects. Consequently, they have applied Language Identification (LID) methodologies to Automatic Dialect Identification (ADI) [19, 28]. Standard LID techniques such as language modeling and phone recognition procedures have shown promising results when treating dialects as subclasses of languages [7, 29]. Dialects, being variations within a common language, generally share vocabulary, syntax, and semantics, with variations primarily occurring in phonology and pronunciation patterns, and minimal differences in grammar. Therefore, methods used for LID may not be directly applicable to ADI [28]. Various studies have focused on dialect processing at different levels, including sentence, word, syllable, or phoneme levels. Both read and spontaneous speech have been considered in text-dependent and text-independent scenarios for dialect processing [18, 22]. Acoustic-phonetic techniques [22, 30, 31] and phonotactic approaches [9, 32, 33] have demonstrated effectiveness in addressing the dialect recognition problem.

Dialect identification can make use of various phonotactic methods when speech transcriptions are accessible, although it can be difficult when transcriptions are unavailable [34]. Within the realm of dialect processing, several techniques receive emphasis, including Parallel Phone Recognition (PPR), Parallel Phone Recognition Language Modeling (PPRLM), and Phone Recognition Language Modeling (PRLM). An example of such a technique is PPRLM, which is employed to distinguish four colloquial dialects from Modern Standard Arabic (MSA) dialects [32, 35].

To characterize dialects and identify their distinguishing features, researchers have explored various phonetic segments. Extensive research has been conducted on the intrinsic characteristics of vowels, including F1, F2, pitch, and duration, to investigate the acoustic distinctions among linguistic dialects [36–39]. For instance, research conducted on the acoustics of Brazilian and European Portuguese dialects revealed substantial disparities in intrinsic vowel attributes, such as F1, F2, pitch, and duration [36]. Another approach has been developed that improves dialect classification performance by fusing phonetic data with vowel acoustic properties [37]. Similar studies have examined how the fifteen normal Dutch vowels are pronounced in different Dutch and Belgian accents by looking at the initial three formants, fundamental frequencies, and duration characteristics [39]. In a recent study focusing on Greek dialects, the dynamic formant frequencies (F1–F4) and acoustic parameters related to vowel duration were taken into account to classify the dialects of Cypriot Greek and Athens Greek [40]. Moreover, it was proposed to compare acoustic and articulatory methodologies to investigate the distinctions in vowels between the English dialects of Australia and America. In this study, vowel data obtained with the aid of electromagnetic articulography were employed, uncovering noteworthy connections between tongue position and formant frequencies [41].

Only a few attempts have been made to differentiate dialects based on vowels in Indian languages. One proposed method involves the utilization of a fuzzy neural network-based system for Assamese dialect recognition. This approach uses prosodic and formant data produced with vowel sounds employing vowel sounds in an acoustic speech stream to identify dialects. They found that the Neural Fuzzy Classifier (NFC) provided a 23% increase in accurate classification rate compared to FFNN, demonstrating its efficacy in

identifying dialects [42]. Koolagudi [43] developed a system capable of recognizing fifteen different languages, including Assamese, Bengali, Hindi, English, and others, by extracting twenty-one Mel-Frequency Cepstral Coefficient (MFCC) features from audio signals obtained from television news networks in India. The extracted features were then used with a GMM-based classifier to differentiate the languages, achieving an average recognition rate of 88% for the 15 languages. Another study by Verma *et al*. [44] presented an automatic language recognition system that used K-means clustering on MFCCs and Support Vector Machines (SVM) for classification. This system achieved an average classification accuracy of 81% for short duration speech signals in English, Hindi, and Tibetan. For dialect identification, Ismail *et al.* [45] worked on Kamrupia and Goalporia dialects as well as the Assamese language using GMM and GMM-UBM techniques. They created a corpus using spontaneous speech and achieved an identification rate of 98.3% with GMM-UBM compared to 85.7% with GMM. Overall, these studies highlight various methods, including neural networks, clustering techniques, and Gaussian mixture models, to effectively recognize and classify dialects based on vowel characteristics in Indian languages. This study focuses on examining the impact acoustic characteristics of vowel on four distinct Hindi dialects. By examining formants (F1, F2 and F3), pitch (F0), and pitch slope data, the researchers explore the eight vowel's acoustic characteristics from Hindi language [46].

Support Vector Machine (SVM) models have proven to be highly effective for prediction and classification tasks, especially when dealing with high-dimensional input spaces. These models are particularly suitable for working with voice representations that contain a large number of features [47]. In a study focused on three Spanish dialects, a hybrid GMM-SVM classifier was utilized to classify the dialects. The experiments encompassed different variables, including formant frequencies, Line Spectral Pairs (LSP), MFCC, intensity, pitch, and zero crossing rate (MEPZ) attributes. Both individual and combined analyses were conducted on these variables [48]. However, training SVMs with a large dataset can lead to increased computational costs. To address this issue, the Minimal Enclosing Ball (MEB) technique is employed as a solution [49].

Chittaragi and Koolagudi [50] used both a single classifier based on the multi-class Support Vector Machine (SVM) technique and a multiple classifier-based ensemble SVM (ESVM) technique to classify Assamese dialects. The ESVM technique has shown superior performance compared to a single SVM. When using spectral features alone, the dialect recognition performance reached 83.12%, while prosodic features achieved a recognition rate of 44.52%. Moreover, the study examines the synergy between spectral and prosodic features by merging their respective feature vectors for dialect recognition. Remarkably, this combination leads to a substantial increase in dialect recognition performance, achieving an impressive 86.25% accuracy. These results suggest the presence of complementary and dialect-specific evidence within both spectral and prosodic features. Notably, when tested on the standard IViE corpus, the ESVM approach achieves a significantly higher recognition rate of 91.38%.

In order to identify Assamese dialects, Das and Bhattacharjee [51] presented a method utilizing the Gaussian Mixture Model (GMM) and Gaussian Mixture Model with Universal Mixture Model (GMM-UBM). By combining MFCC and ΔMFCC features and applying this model, they achieved an identification accuracy of 97.57%.

Sarmah and Dihingia [52] utilized a random forest approach to identify Assamese dialects based on acoustic features of Assamese vowels, achieving a classification accuracy of 94.0% for the test data.

In general, for classification statistical i.e. probabilistic or rule based approach were used by single classifier. The single classifier algorithm depends on a single approach to achieve classification performance. The classification method used Gaussian Mixture Model (GMM), Linear Discriminate Analysis (LDA), Support Vector Machines (SVM) and neural network [10, 24, 53, 54].

Recently, the concept of combining multiple classifiers has gained attention as a means to enhance performance compared to using individual classifiers alone. One approach to tackle robustness issues across various dialects is the utilization of rotation forest, which is an ensemble of decision trees [22]. However, there have been limited efforts to apply ensemble approaches for addressing dialect identification challenges [30, 55]. Similarly, the AdaBoost ensemble method has been employed for word-based dialect detection, where instead of operating in the feature space, this approach focuses on improving performance [30]. These ensemble techniques have demonstrated significant improvements compared to using single classifiers. Furthermore, the majority of studies in this domain have primarily relied on n-gram features in natural language processing to identify dialects from datasets that are text-based [56].

Nowadays, the majority of electronic devices rely on automatic speech recognition systems. For dialect recognition to be efficient, especially in resource-constrained scenarios, it is crucial to have faster and simpler computations. Previous research indicates that most existing systems consider dialect recognition when processing longer-duration data, such as the complete signal. Nevertheless, these existing models suffer from computational complexity issues and lack language independence. When applied to another language, their performance may not be as reliable, given the significant variations in pronunciation patterns among different languages.

The absence of a standard Automatic Dialect Identification (ADI) system for the Assamese language serves as the driving force behind the development of a novel system that aims to characterize and identify the four dialects of the Assamese language.

## III. Experimental Setup

### A. Creating a Vowel Database and Assembling a Speech Database for Assamese Dialects

The vowel database utilized in this work was generated by gathering spoken samples from indigenous speakers of four distinct Assamese dialects, encompassing read and semi-spontaneous speech in both. The recording sessions primarily involved participants from rural areas who were either native-born or had lived there for a considerable duration. Most of the speakers had a minimum educational qualification at or below the level of matriculation, indicating a lower level of formal education. Their way of speaking in their native dialect is less influenced by written and standard Assamese due to their limited education. Approximately 90% of the selected speakers only spoke Assamese. The speakers' ages ranged from 25 to 65 years old.

The recordings were made with a Sony voice recorder that has a 44.1 kHz sampling rate and a 16-bit mono resolution for every sample. The speech data consists of 10 individuals, five male and five female, who can read and speak semi-spontaneous styles and represent all of the Assamese dialects. Separate database was created for the read speech and the semi-spontaneous speech. The recording environment was relatively quiet. For the read speech dataset, a script with a rich set of phonetic content was used. For instance, semi-spontaneous speech, random questions were asked to elicit natural and spontaneous conversations about topics such as childhood, schooling, personal history, and professional experience. The recorded data was subsequently subjected to pre-processing to eliminate noise and prominent pauses in the speaker's speech. Careful selection of speakers was done based on the collected database, ensuring that their speech was clear and understandable, to create the vowel database. Table II presents a concise summary of the primary Assamese dialect database, which served as the foundation for generating the vowel database.

The process illustrated in Fig. 2 demonstrates the manual segmentation technique for identifying vowels using the Praat open-source software tool [57]. To create the vowel database, a minimum of ten speakers with clear speech from the actual Assamese dialect database are selected. The manual segmentation focuses on breaking down continuous speech into individual words. Specifically, words that exhibit well-articulated and interesting vowels are chosen. Words with phonetic unit patterns like /VcV/ and /cVcV/—where V stands for a vowel and C for a consonant—are favored when vowels are retrieved from spontaneous speech. This selection is made considering that the co-articulation of the particular consonants has a relatively negligible impact on the stability of the vowels preceding and following them. Vowels are identified from the particular phonetic unit by visually inspecting the waveform, formants, and intensity parameters in Praat. When there is a progressive increase in intensity, the F1 and F2 values are examined to determine the onset of the vowel. The F1 and F2 readings at the beginning of a progressive reduction in intensity might also be used to predict the vowel's termination. The segment between the beginning (onset) and end (offset) of the vowel, known as the steady-state portion, is regarded as the signal that represents both short and long vowels. Since, the majority of the vowels are derived from the beginning and end of words because co-articulation has a lesser impact in these portions.

TABLE II. Assamese Dialect Vowel Database Information

| Dialects | Total Speaker | Read Mode (in Minutes) | Semi-read Mode (in Minutes) | No. of Vowel (In Total) |
|---|---|---|---|---|
| **Eastern** | 10(5M+5F) | 58 | 54 | 540 |
| **Central** | 10(5M+5F) | 63 | 58 | 540 |
| **Kamrupia** | 10(5M+5F) | 58 | 56 | 540 |
| **Goalporia** | 10(5M+5F) | 63 | 58 | 540 |

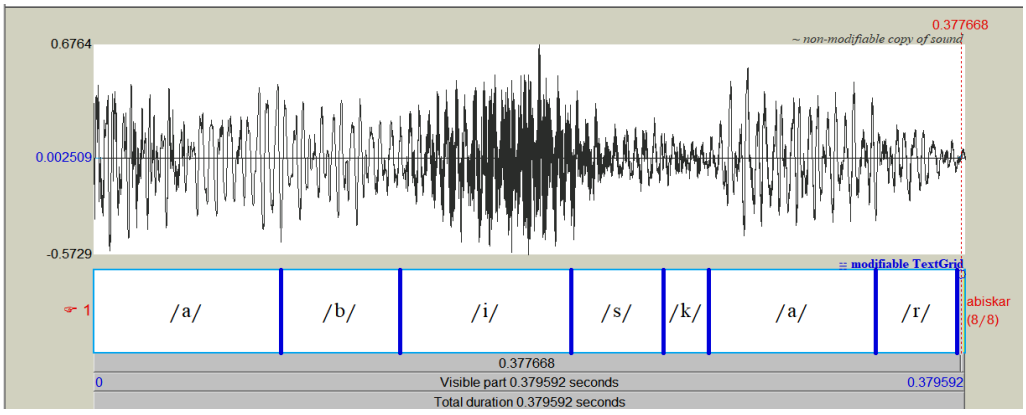


Fig. 2. Manual transcription of the word "Abiskar" (Assamese word for Discover) in .wav file using Praat.

### B. Feature Extraction

This research advances the ADI system by employing acoustic-phonetic information extracted from Assamese vowels. The acoustic properties, such as formant frequencies, obtained from the steady-state regions of the vowels play an important role in distinguishing between different vowel sounds. Additionally, the statistical analysis of each feature derived from the vowels gives

details about their relevance in the classification of dialects based on vowel sounds. It is crucial to consider factors like gender differences, emotional states, co-articulation effects, the setting of sound units, and articulatory configurations, as they contribute to the acoustic variation observed in vowel sounds [58]. Fig. 3 displays the mean LP spectra of the vowel /e/ as pronounce by a speaker randomly chosen from the four Assamese dialects. The image demonstrates the presence of distinct spectral shapes among the dialects, indicating systematic and significant differences in vowel intrinsic spectral features. The variations in energy levels, spectral peaks, spectral sharpness, and the positions of formant frequency values (F1–F4) can be discerned in these differences. In this study, the formant frequencies F1 to F3, along with pitch, speech frame energy, and vowel duration, are employed to characterize the four dialects. Formants, which are resonant frequencies in connection with shape of the mouth cavity during phoneme production, play an important part in vowel recognition. Among the formants, F1 is associated with vowel height, where low values correspond to high front vowels (/i/, /e/, /u/, and /u:/ as shown in Table I), while high values are linked to low mid vowels (/a/ and /a:/). Similarly, F2 is related to tongue advancement, with high values indicating fronting and low values indicating backing. F3 is commonly used to distinguish between rounded and unrounded vowels and the degree of lip rounding and constriction influences F2 and F3 proportionally.

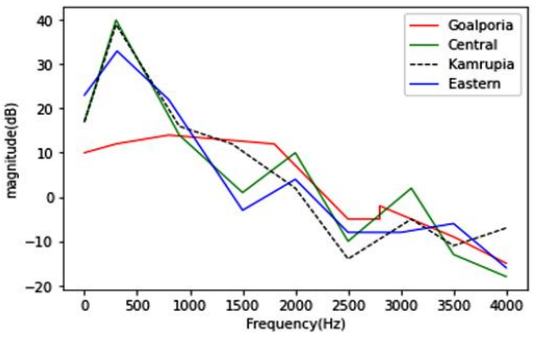


Fig. 3. Typical LP spectrum of vowel /e/ in four dialects.

Instead of relying solely on energy attributes, it is advantageous to analyze the dynamic and static behaviors of the steady-state region of a vowel to characterize languages. The process employed in this research involves retrieving three formants (F1, F2, F3), pitch and energy contours, along with local variations in vowel pronunciation across dialects. To represent the temporal dynamics of phonation, intonation, and loudness in local variations, six features are extracted from formants, pitch, and energy using a Legendre polynomial fit function of order five. To create a feature vector with a size of 5×6, i.e., 30, these characteristics are joined in the following order: F1, F2, F3, pitch, and energy. Additionally, the duration of each individual vowel in milliseconds is incorporated into the feature vector to capture regional variations, resulting in a feature vector of dimension 25. Vowels are used to extract both regional and global properties. The mean, minimum, maximum, standard deviation and variance of F1, F2, F3, pitch, and energy are retrieved for each vowel. These represent the static characteristics of the vowels. Consequently, four static features are derived from each vowel, encompassing F1, F2, F3, pitch, and energy, resulting in a feature vector of 25 dimensions.

### C. Statistical Evaluation of Features

A conventional F2−F1 plot was created using the mean F2 and F1 values of all five vowels from four dialects, as shown in Fig. 4. The plot arranges the vowels in the order of fronting to backing: /i/, /e/, /a/, /o/, and /u/. The Central and Kamrupia dialects can be distinguished from other dialects by their higher F1 and F2 values for front vowels. These dialects exhibit distinct speech styles characterized by stronger energy levels, higher vocal tonalities, and faster speech rates, which contribute to larger F2 values for front vowels. All dialects, however, share almost identical F1 and F2 values for the central-low vowel /a/. In the Goalporia dialect spoken in lower Assam, the vowels /e/, /a/, and /u/ have lower F1 and F2 values, while the back-mid vowel /o/ exhibits higher F1 and F2 values. Speakers from this region tend to exhibit less variation from the standard Assamese pronunciation when they are conscious. Similarly, the Eastern dialect also shows small F1 and F2 values for its vowels.

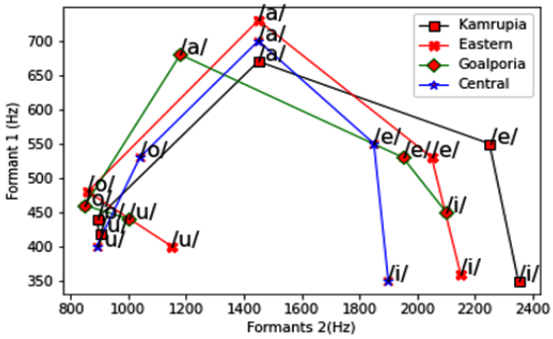


Fig. 4. Vowel articulation in four distinct dialects, with F2 and F1.

The Eastern and Central speaking styles resemble the written/standard form of Assamese more closely, with each phonetic phrase being distinct. The plot Fig. 4 clearly demonstrates the presence of different vowel pronunciation patterns among the four dialects. Fig. 5 illustrates the F1, F2, and F3 formants for the vowels /a/, /i/, and /u/ in Assamese dialects, obtained through Legendre curve fitting of order five using male speakers. The figure clearly demonstrates significant variations in the vowels across different dialects. These three vowels, due to their distinctiveness, are considered as representative formants in the plot. For instance, the vowel /i/ is characterized by a lower F1 and a higher F2, while the vowel /u/ exhibits a lower F1 and a higher F2. In contrast, the vowel /a/ displays a higher F1 and a lower F2 [59].

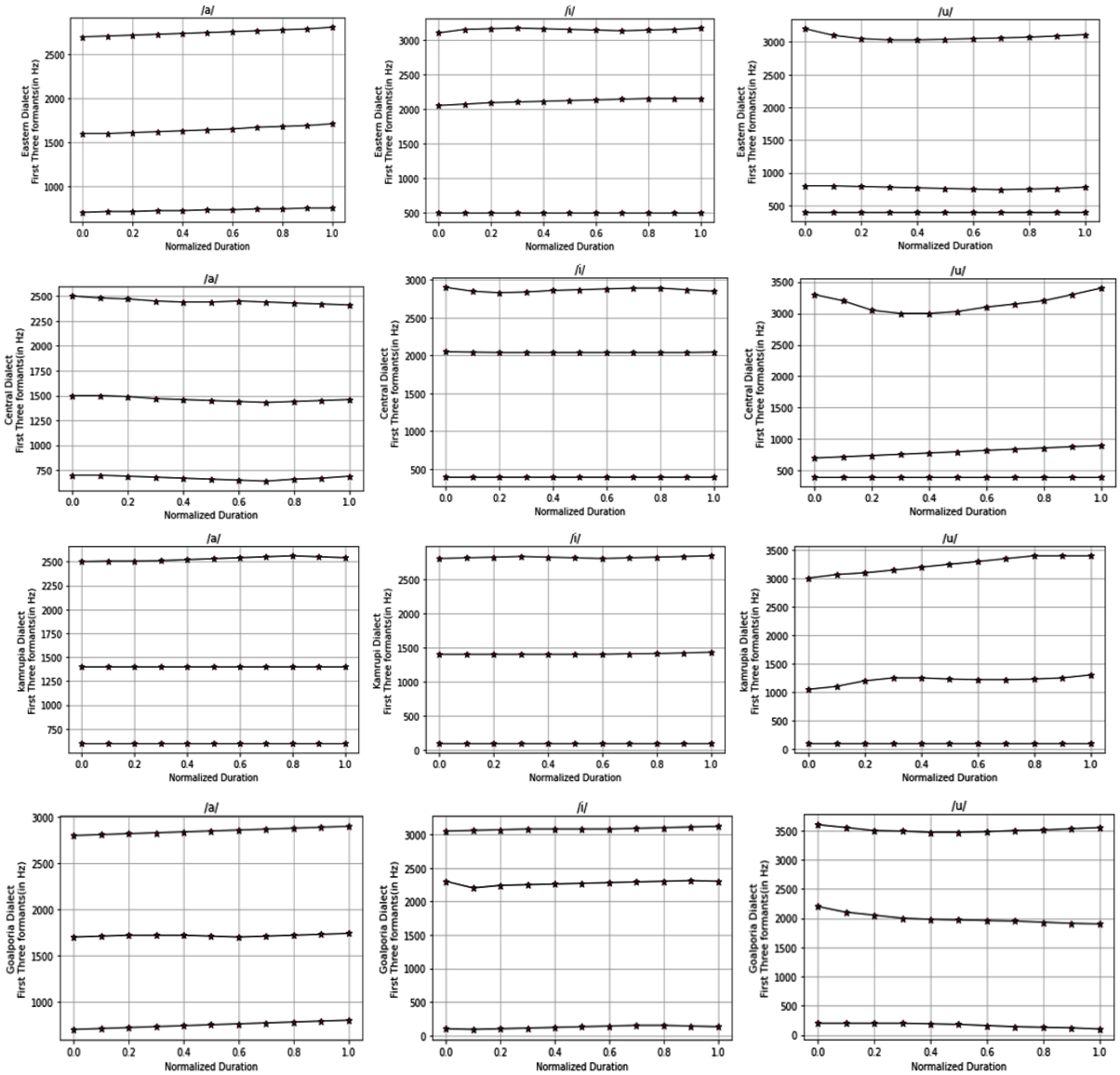


Fig. 5. Three formants (F1, F2, F3) for three vowels (/a/, /i/, /u/) for all four dialects.

To assess the impact of speakers' native dialect, ANOVA tests were conducted on formant frequencies, pitch slope, intensity, and duration. For the purpose of identifying the acoustic factors associated with these dialects, an ANOVA test is executed on the vowel sound units in Assamese. Agrawal *et al*. [60] examine the variations in Hindi dialects by employing Analysis of Variance (ANOVA) on acoustic attributes including formant frequency, pitch, pitch slope, duration, and intensity in vowel sounds. They employ a kernel-based Support Vector Machine (SVM) to gauge the capacity of these acoustic features to distinguish dialects, achieving a classification accuracy of 66.97%. Through the incorporation of shifted delta cepstral coefficients with Mel-Frequency Cepstral Coefficients (MFCC), the accuracy rises to 74% for prosodic feature combinations, and when spectral and prosodic features are combined, they attain a classification accuracy of 88.77%. One-way ANOVA is used for the statistical examination of the features. The mean values of F1, F2, F3, and duration features for all are computed of the eight vowels. Tables III and IV present the outcomes of the F-test, including the F-statistic, mean values, and standard deviations of the eight vowels across the four different dialects. The F-statistic, or F-ratio, determines whether the means of different samples significantly differ. However, relying solely on the F-statistic is often insufficient, so the *P*-value is also considered. The *P*-value indicates the likelihood of obtaining the observed results. In this analysis, the hypothesis is that the formants extracted from the vowels exhibit significant differences among the four dialects. The significance threshold α is set at 0.05, meaning that if the *P*-value is below 0.05, the observed differences are considered statistically significant.

Table III reveals that the Kamrupia dialect exhibits higher F1 values for the /i/, /a/, /o/, and /o:/ vowels. In the Goalporia region, front vowels (/i/, /e/, /e:/, and /a/) are associated with high F2 values. The Eastern, Central, and Kamrupia dialects share a common characteristic of

having low F2 and F3 values across all vowels. The Central region shows higher F3 values for the /e:/, /a/, and /u/ vowels, while the Eastern region exhibits higher F3 values for the /i/, /i:/, /e/, and /u:/ vowels. The *P*-values for the /e/ and /u:/ vowels, highlighted in bold in Tables III and IV, are found to be less than 0.05. Furthermore, the f-statistic is greater than the critical f-value, indicating significant pronunciation variations among dialects. For example, the ANOVA results for the /e/ vowel are as follows: [$f(4,434) = 4.51$, $P = 0.0029$] for F1, [$f(4,434) = 2.71$, $P = 0.034$] for F2, and [$f(4,434) = 3.65$, $P = 0.017$] for F3. Here, $f(4,434)$ denotes the f-statistic with degrees of freedom representing the within-group (N–k) and between-group (number of classes –1, 4–1). Similarly, it has been noted that the /u:/ vowel varies significantly across dialects with [$f(4,220) = 10.34$, $P = 0.0001$] for F1 and [$f(4,220) = 10.05$, $P = 0.0001$] for F2 values. However, F3 does not contribute significantly to the differentiation of dialects containing the /u:/ vowel.

Table III represents the F2 statistics, and it is noted that the vowel /o/ shows [$f(4,220) = 9.19$, $P = 0.0006$], indicating significant differences in F2 values among the dialects. Furthermore, variations in F2 values are noted across the dialects. Moreover, statistically significant differences in F2 values among the four Assamese dialects were found for the vowels /i:/ [$f(4,210) = 4.90$, $P = 0.002$] and /e/. These tables demonstrate that only a subset of vowels exhibit a range of F1, F2, and F3 values. These characteristics (F1, F2, F3) can be leveraged for automatic classification as they exhibit substantial changes across the four Assamese dialects, providing strong evidence for differentiation.

Table IV presents the results of an ANOVA test conducted to analyze the statistical significance of vowel duration in differentiating the four dialects. It was hypothesized that vowel duration plays a crucial role in distinguishing between the dialects. The outcomes show significant differences for the vowels /i/ [$f(4,210) = 2.67$, $P = 0.041$], /a/ [$f(4,310) = 6.01$, $P = 0.0002$], /a:/ [$f(4,160) = 4.31$, $P = 0.0053$], /o/ [$f(4,200) = 9.1$, $P = 0.0005$], /u/ [$f(4,240) = 4.91$, $P = 0.005$], and /u:/ [$f(4,200) = 7.41$, $P = 0.0005$]. The Kamrupia dialect typically has shorter vowel durations. With the exception of /e/, most vowels demonstrate differences among the dialects.

TABLE III. COMPUTE MEAN AND STANDARD DEVIATION OF THE F1 FEATURE FOR EIGHT LONG AND SHORT VOWELS USED IN THE F-TEST, SD (STANDARD DEVIATION), EASTERN (L-1), CENTRAL (L-2), KAMRUPIA (L-3), AND GOALPORIA (L-4) PAPERS

| Assamese vowels | F1 formant | | | F2 formant | | | F3 formant | | |
|---|---|---|---|---|---|---|---|---|---|
| | f-stat | *P*-value | F-crit | f-stat | *P*-value | F-crit | f-stat | *P*-value | F-crit |
| **/i/** | 0.51 | 0.68 | 2.59 | 1.25 | 0.27 | 2.55 | 1.77 | 0.145 | 2.56 |
| **/i:/** | 2.01 | 0.12 | 2.65 | 1.34 | 0.25 | 2.63 | 4.90 | **<0.002** | 2.65 |
| **/e/** | 4.51 | **<0.0029** | 2.51 | 2.71 | **<0.034** | 2.54 | 3.65 | **<0.017** | 2.53 |
| **/e:/** | 0.94 | 0.42 | 2.58 | 0.97 | 0.39 | 2.58 | 1.02 | 0.387 | 2.59 |
| **/a/** | 1.85 | 0.09 | 2.39 | 1.98 | 0.08 | 2.41 | 1.04 | 0.372 | 2.39 |
| **/a:/** | 1.41 | 0.21 | 2.60 | 0.33 | 0.81 | 2.61 | 1.44 | 0.221 | 2.61 |
| **/o/** | 2.29 | 0.10 | 3.09 | 9.19 | **<0.0006** | 3.06 | 1.24 | 0.324 | 3.09 |
| **/o:/** | 0.86 | 0.47 | 3.01 | 1.11 | 0.37 | 3.06 | 1.41 | 0.265 | 3.01 |
| **/u/** | 1.20 | 0.31 | 2.81 | 2.14 | 0.09 | 2.89 | 0.60 | 0.654 | 2.98 |
| **/u:/** | 10.34 | **<0.0001** | 2.86 | 10.05 | **<0.0001** | 2.85 | 0.71 | 0.574 | 2.88 |

TABLE IV. COMPUTE MEAN AND STANDARD DEVIATION OF DURATION IN MILLISECONDS AND PERFORM AN F-TEST FOR 8 LONG AND SHORT VOWELS, SD-STANDARD DEVIATION, EASTERN (L-1), CENTRAL (L-2), KAMRUPIA (L-3), AND GOALPORIA (L-4)

| Vowels | Duration(ms)-mean | | | | Duration(ms)-SD | | | | F-Test | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L4 | L1 | L2 | L3 | L4 | f-stat | *P*-value | F-crit |
| **/i/** | 60 | 59 | 52 | 75 | 12 | 11 | 3 | 14 | 2.65 | **<0.041** | 2.54 |
| **/i:/** | 105 | 96 | 87 | 100 | 9 | 18 | 18 | 20 | 0.97 | 0.418 | 2.64 |
| **/e/** | 56 | 68 | 54 | 62 | 10 | 9 | 9 | 13 | 1.96 | 0.112 | 2.54 |
| **/e:/** | 106 | 124 | 116 | 112 | 24 | 26 | 8 | 17 | 1.06 | 0.373 | 2.61 |
| **/a/** | 55 | 61 | 51 | 61 | 11 | 14 | 8 | 14 | 6.01 | **<0.0002** | 2.41 |
| **/a:/** | 111 | 105 | 87 | 103 | 20 | 13 | 6 | 7 | 4.32 | **<0.0053** | 2.60 |
| **/o/** | 78 | 63 | 49 | 73 | 11 | 13 | 2 | 2 | 9.1 | **<0.0005** | 3.9 |
| **/o:/** | 114 | 104 | 90 | 115 | 5 | 5 | 11 | 10 | 2.59 | 0.079 | 3.9 |
| **/u/** | 56 | 93 | 56 | 63 | 9 | 26 | 6 | 11 | 4.93 | **<0.005** | 2.81 |
| **/u:/** | 123 | 128 | 83 | 117 | 24 | 12 | 6 | 8 | 7.41 | **<0.0005** | 2.79 |

The outcomes of the ANOVA test carried out on the eight vowel durations distinctly demonstrated that, in the case of the majority of the vowels, one set of speakers differed from another group of speakers. It can be deduced that the duration of most vowels can serve as a discriminative factor for identifying dialects.

### *D. Prosodic Features*

Features extracted from the frame of the signal only capture limited local information. As a means to capture changes within and across sequences of sound units, it is necessary to derive some features from a larger time frame of the speech. Elements such as intonation, intensity patterns, and varying speaking rates contribute to the naturalness of conversational speech [61]. These aspects, known as prosodic cues, facilitate the utilization of distinct speaking patterns unique to each dialect. Prosodic features are typically examined through pitch, intensity fluctuations, stress patterns, and rhythmic output. These additional characteristics of speech units offer valuable insights for dialect identification purposes.

Prominent distinctions in acoustic and linguistic characteristics have been discovered between spontaneous and read speech, even when utilizing two distinct speech datasets from the Assamese dialect speech corpus. These differences encompass intonation, loudness, speaking rate, and other perceptual aspects of speech. By incorporating these features, speech acquires a more natural quality by the utilization of intonation variations, varied durations, and intensity patterns. Dialect identification primarily relies on discerning distinct pronunciation patterns employed by different speakers within the same language. It has been established that prosodic variations genuinely exist in speech and play a crucial part in conveying dialect-specific information [22]. Reports indicate that prosodic Variations frequently occur in a majority of Indian dialects [8, 24, 62].

To detect prosody cues, pitch, energy, and duration parameters are derived from shorter segments of vowels. A pitch estimation technique depending on the subharmonic-to-harmonic ratio is used to extract pitch (F0) information [63]. Intonation patterns, which encompass the differences in rise and fall of pitch over time, help identify specific dialectal patterns. The energy level of the speech signal is utilized to calculate the voiced and unvoiced parts of speech. By combining energy, pitch, and duration, the stress patterns of speakers can be expressed. Frame energy, a prosodic property frequently employed in accent and dialect recognition research, is considered by some researchers as a separate stream [16], while others incorporate it alongside spectral features [64]. Both approaches improve the system's efficiency. The energy of each segmented speech frame, which overlaps with adjacent frames, is computed by summing the squared amplitudes of each sample.

The speech signal exhibits time-varying energy characteristics, which might be thought of as representing the loudness property. Short-time energy, derived from vowel sounds, is used to capture this aspect of loudness, which plays an important role in how sound is perceived by individuals. The development of energy over time is assessed by comparing the variations in sample amplitudes within a frame. Eq. (1) is utilized to compute the short-term energy feature.

$$E(i) = \sum_{n=1}^{W_L} |x_i(n)|^2 \quad (1)$$

Here, $x_i$ $(n)$, $n=1, 2...$ $W_L$ is the audio samples in the frame with $W_L$ denoting the frame's length.

$$E(i) = \frac{1}{W_L} \sum_{n=1}^{W_L} |x_i(n)|^2 \quad (2)$$

To eliminate the dependence on the frame length, energy is normalized over a frame by dividing it with $W_L$.

A box plot employed to represent the first-order statistics of energy values taken from voice samples among the four Assamese dialects. Fig. 6 presents the statistical information, including the median, maximum, minimum, and first and third quartiles. It is obvious that the Kamrupia dialect exhibits a larger interquartile range in relation energy. This suggests that speakers of this dialect have higher overall energy levels and a greater variability in their energy patterns. In contrast, the Central and Eastern dialects have a narrower range of energy values, indicating lower energy levels and a more consistent speaking pattern across the region.

Goalporia dialect, similar to Kamrupia, shows a wider energy range because of the usage of higher energy values in speech. The interquartile ranges of the Central and Eastern dialects are observed to be similar. Outliers, denoted by the symbol "+", are present in all dialects except for Goalporia, and they are usually skewed towards the maximum values The findings from Figs. 4–6 represent that the Kamrupia and Goalporia dialects exhibit higher F1 and F2 values, which are connected to a wider interquartile range in those dialects. Moreover, the energy feature shows higher values in these dialects as well. Conversely, the central and eastern dialect region is related to lower F1 and F2 values, together with a noticeably narrower interquartile range. Interestingly, there is correlation observed between F1, F2 values, and the energy features.

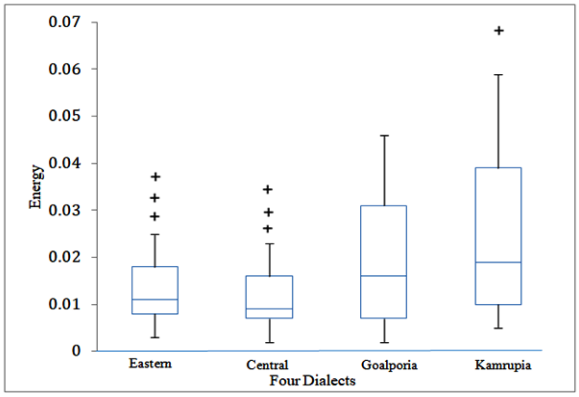


Fig. 6. Energy data for speakers of four dialects.

## E. Prosodic Features

This study implements an approach for dialect identification that combines acoustic and prosodic behaviors. The main objective is to look at the performance of individual classifiers and ensemble classifiers (Random Forest, Extreme Random Forest, and Extreme Gradient Boosting) in dialect identification using acoustic and prosodic data. Individual classifiers use statistical techniques to estimate class-conditional probability, while ensemble classifiers aggregate the predictions of multiple base models to improve accuracy. Ensemble classifiers aim to leverage the collective wisdom of a panel of experts rather than relying solely on the judgment of a single expert (base learner). The selection of base models can be done through independent or dependent approaches. Bagging methods are employed to combine predictions from different base models obtained from bootstrap samples of the initial data. Boosting algorithms, on the other hand, grow the base models in a dependent manner, iteratively modifying them based on training to reduce errors in the ensemble. The workflow of both single and ensemble classification algorithms is illustrated in Fig. 7.

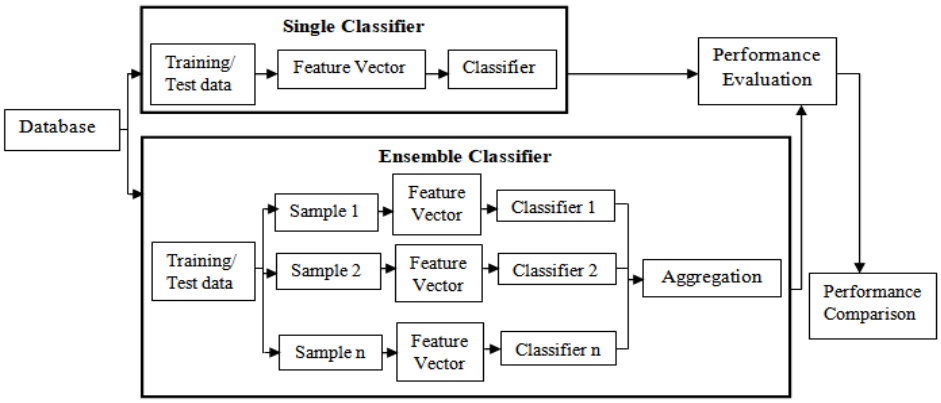


Fig. 7. Single vs. ensemble classifier workflow.

Ensemble methods have emerged as highly effective strategies for voice recognition tasks by combining predictions from multiple classifiers. These methods leverage the strengths of different algorithms, thereby enhancing the overall predictive performance. In this study, dialect identification algorithms are developed using both boosting, specifically Extreme Gradient Boosting (XGB), and bagging techniques, including Random Forest (RF) and Extreme Random Forest (ERF).

During the second phase of the analysis, the individual models within the ensemble are compared to one another [48, 65]. This study introduces a Random Forest (RF) classifier built on decision trees. A forest comprising 2048 decision trees is created using a combination of randomized tree predictors and bootstrapping techniques applied to the training dataset.

Through the utilization of 2048 decision trees and the Assamese vowel dialect speech corpus, empirical analysis yields higher accuracy results. During the construction of each tree, the split-node procedure is regulated by selecting the ideal split in light of the Gini criterion selected at random from a set of features.

$$Gini = N_L \sum_{t=1\ldots T} p_{kL}(1 - p_{kL}) + N_R \sum_{t=1\ldots T} p_{kR}(1 - p_{kR}) \quad (3)$$

The quantities $p_{kL}$ and $p_{kR}$ indicate the ratios of the $t$ class in the left side and right side nodes of the tree

The quantities $p_{kL}$ and $p_{kR}$ indicate the ratios of the $t$ class in the left side and right side nodes of the tree, while $N_L$ and $N_R$ represent the respective counts of nodes in the left side and right side of the tree. The division of a node involves considering $\checkmark n$ features, where $n$ is equivalent to the length of the feature vector. In the case of constructing the entire forest using decision trees, the categorization process involves aggregating the predictions made by multiple trees trained on different subsets from the training set through voting.

An ERF, that is a simplified version of RF, constructs 2048 randomized trees by sampling with replacement. Unlike RF, ERF selects the optimal threshold for every potential feature from a set of randomly generated thresholds, rather than using an optimized split. Similar to RF, the maximum features parameter in ERF is also set to $\checkmark n$, where n represents the size of the feature vector.

In XGB Boosting, the forecasts of the base learner are progressively improved in a greedy manner, aiming to reduce the selected loss function (error) and enhance accuracy. In this study, the multi-class logloss function is employed as the chosen loss function for XGB Boosting.

$$logloss = -\frac{1}{p}\sum_{r=1}^{P}\sum_{s=1}^{Q} z_{i,j} \log\left(t_{i,j}\right) \quad (4)$$

where $P$ represents the dimension of the feature vector, $Q$ denotes the total count of class labels, and $Z_{i,j}$ of the base learner takes the value 1 when observation $r$ corresponds to class $s$, and $0$ otherwise. The projected probability of observation $r$ belonging to class $s$ is denoted as $t_{i,j}$

In this study, a decision tree classifier is used as the base learner. The decision tree construction involves several steps: setting the learning rate (η) to 0.2 to control the reduction of feature weights and make the boosting process more cautious, imposing a maximum depth limit of 6 for each tree, and using a subsample ratio of 0.6 for training data instances. The objective function SoftMax is employed to handle the four classes in dialect recognition. These parameter values are chosen empirically to optimize the recognition accuracy. The XGBoost library is utilized for implementing the system [66, 67]. The proposed dialect recognition system incorporates three decision tree and ensemble approaches, and its block design is depicted in Fig. 8.

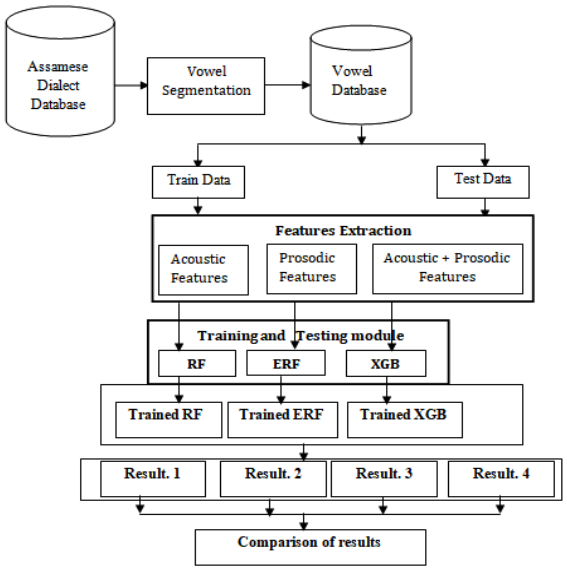


Fig. 8. Dialect recognition system using ensemble.

## IV. Result and Discussion

To classify dialects based on their static and dynamic characteristics, acoustic-phonetic and prosodic features are employed in experimental settings. The aim is to develop Automatic Dialect Identification (ADI) systems that utilize both static and dynamic feature vectors. Three decision tree base learners and ensemble methods are used in these systems. Test speech samples are provided to all ADI systems, and the dialect corresponding to the speech sample is assumed being the model with the strongest supporting evidence. The efficiency of dialect classification is assessed by using two validation techniques. One is Simple Validation (SV) and the other one is Cross-fold Validation (CV). Both individual and combined features are utilized to evaluate the effectiveness of integrating acoustic and prosodic characteristics in identifying dialects.

Cross-validation is utilized to address the dataset's variations and improve the model's stability. This method entails splitting the data into several subsets, enabling the training and testing to be performed on various combinations of the data. This research utilizes a five-fold cross-validation strategy, where four sets are employed for training purposes, while one set is designated for testing. In each iteration, the training and testing folds are interchanged to consistently evaluate the performance of the system. The database is split into an 80:20 ratio, 80% data is used for training (1728 vowels) and 20% used for testing (432 vowels). The identical procedure is replicated five times, employing distinct combinations of the 80% training data each time. The final predictions for the 20% testing results are determined by aggregating the majority votes from each of the five predictions. This approach provides a more robust and accurate prediction of the model's performance.

Table V displays the performance outcomes of the models on static and dynamic features, including Precision, Recall, and F1-score. Additionally, it presents the macro average across all dialects. Regarding static features, our findings demonstrate that ERF achieves the highest performance with an average F1-score of 0.756, followed by XGB, which also achieves an average F1-score of 0.643. Regarding the performance of individual dialects, it is evident that all models achieved their most impressive results for the dialect variations, particularly Goalporia, where the ERF model attained an outstanding F1-score of 0.806. Table VI displays the outcomes of all models when utilizing dynamic features. The three models used in the evaluation are the same as before. Among these models, the ERF model achieved the highest average score, obtaining an average F1 score of 0.825.

The effectiveness of the vowel-based ADI system, which focuses on dynamic behavior, is summarized in Table VII. The results presented in the table consider only the Cross-Validation (CV) outcomes for further analysis. The table demonstrates the individual contributions of each formant frequency and prosodic characteristic of vowels to the overall performance. When dynamic formant frequencies are utilized for classifying Assamese dialects, the system achieves a performance of 61.30%. Similarly, when considering the vowel regions and their dynamic features such as F0, energy, and duration, the system achieves an identification accuracy of approximately 58.10%. To leverage the complementary information provided by both features, they are combined into a new feature vector. By using this merged feature vector, the result of the system improves significantly to 69.51%. This indicates the effectiveness of integrating both prosodic patterns and formant frequencies for dialect recognition.

TABLE V. The Performance Scores of Individual Dialect Varieties Using Three Models: RF, ERF, and XGB, with Static Features. the Evaluation Metrics Used Are Recall, Precision, and F1-score. The Macro Average is Reported to Provide an Overall Average

| Dialects | Model | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Eastern | RF | 0.772 | 0.721 | 0.746 |
| | ERF | 0.755 | 0.744 | 0.749 |
| | XGB | 0.754 | 0.705 | 0.729 |
| Central | RF | 0.731 | 0.851 | 0.786 |
| | ERF | 0.722 | 0.765 | 0.742 |
| | XGB | 0.750 | 0.848 | 0.769 |
| Kamrupia | RF | 0.533 | 0.490 | 0.485 |
| | ERF | 0.728 | 0.729 | 0.729 |
| | XGB | 0.567 | 0.446 | 0.525 |
| Goalporia | RF | 0.579 | 0.527 | 0.512 |
| | ERF | 0.833 | 0.782 | 0.806 |
| | XGB | 0.581 | 0.458 | 0.552 |
| Macro | RF | 0.653 | 0.642 | 0.632 |
| | ERF | 0.759 | 0.755 | 0.756 |
| | XGB | 0.663 | 0.619 | 0.643 |

TABLE VI. The Performance Scores of Individual Dialect Varieties Using Three Models: RF, ERF, and XGB, with Dynamic Features. The Evaluation Metrics Used Are Recall, Precision, and F1-Score. The Macro Average is Reported to Provide an Overall Average

| Dialects | Model | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Eastern | RF | 0.705 | 0.885 | 0.784 |
| | ERF | 0.714 | 0.900 | 0.796 |
| | XGB | 0.710 | 0.890 | 0.789 |
| Central | RF | 0.677 | 0.882 | 0.766 |
| | ERF | 0.896 | 0.832 | 0.862 |
| | XGB | 0.722 | 0.787 | 0.753 |
| Kamrupia | RF | 0.698 | 0.787 | 0.739 |
| | ERF | 0.900 | 0.753 | 0.819 |
| | XGB | 0.776 | 0.725 | 0.749 |
| Goalporia | RF | 0.645 | 0.724 | 0.682 |
| | ERF | 0.767 | 0.890 | 0.823 |
| | XGB | 0.689 | 0.756 | 0.720 |
| Macro | RF | 0.681 | 0.819 | 0.742 |
| | ERF | 0.819 | 0.843 | 0.825 |
| | XGB | 0.724 | 0.789 | 0.752 |

The ADI system incorporates global features, including statistical mean values, along with its analysis. Table VIII presents an overview of the efficiency of the vowel-based ADI system, focusing on static behaviors. The highest performance in dialect recognition is achieved using the ERF model, which utilizes both acoustic and prosodic characteristics, with recognition scores of approximately 63.44% and 65.96%, respectively. Vowel prosody traits contribute to a slight improvement in the categorization rate. Just like the

results observed for dynamic features, the utilization of the combined feature vector leads to higher rates of dialect recognition. Specifically, the CV (cross-validation) settings yield a recognition rate of 76.84%, while the SV (simple validation) settings achieve a recognition rate of 79.89%. Fig. 9 provides a comparison of the categorization outcomes obtained from the three classification methods.

TABLE VII. PERFORMANCE OF DIALECT RECOGNITION WITH DYNAMIC FEATURES VECTORS

| Features | Recognizability rate in % | | | | | |
|---|---|---|---|---|---|---|
| | RF | | ERF | | XGB | |
| | SV | CV | SV | CV | SV | CV |
| **Formants (Acoustic)** | 62.83 | 57.55 | 63.30 | 61.30 | 64.88 | 56.97 |
| **F0+E+Dur.(Prosodic)** | 60.82 | 55.55 | 61.65 | 58.10 | 56.08 | 55.86 |
| **Formants+F0+E+Dur.(Acoustic+prosodic)** | 70.26 | 65.47 | 73.88 | 69.51 | 72.94 | 64.95 |

TABLE VIII. PERFORMANCE OF DIALECT RECOGNITION WITH STATIC FEATURES VECTORS

| Features | Recognizability rate in % | | | | | |
|---|---|---|---|---|---|---|
| | RF | | ERF | | XGB | |
| | S-V | C-V | S-V | C-V | S-V | C-V |
| **Formants(Acoustic)** | 69.89 | 62.33 | 69.51 | 63.44 | 69.89 | 58.33 |
| **F0+E+Dur.(Prosodic)** | 71.66 | 66.65 | 72.21 | 65.96 | 66.65 | 59.80 |
| **Formants+F0+E+Dur.(Acoustic+prosodic)** | 78.32 | 75.23 | 79.89 | 76.84 | 76.66 | 74.31 |

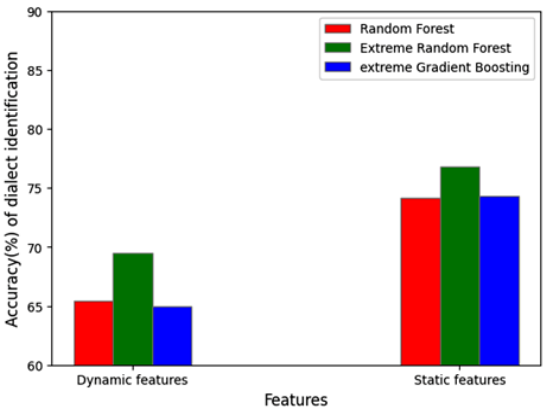


Fig. 9. Evaluation of dynamic and static features' performance.

Fig. 10 displays confusion matrices that illustrate the performance of the ADI system when employing both dynamic and static features. These matrices provide a comprehensive understanding of the recognition results achieved by combining formants with prosodic characteristics without sacrificing generality. Fig. 10(a) displays the confusion matrix for the ERF model, which achieves a typical precision of 69.51% approximately. Similarly, Fig. 10(b) shows the confusion matrix employing global features, leading to average recognition accuracy 76.84% in identifying dialect. Confusion matrices are shown in Fig. 10 to shows how the ADI system performs when dynamic and static features are used in both. These matrices provide a comprehensive understanding of the recognition results achieved by combining formants with prosodic characteristics without sacrificing generality. Fig. 10(a) displays the confusion matrix for the ERF model, which achieves a typical precision of 69.51% approximately. Similarly, Fig. 10(b) shows the confusion matrix employing global features, leading to average recognition accuracy 76.84% in identifying dialect.

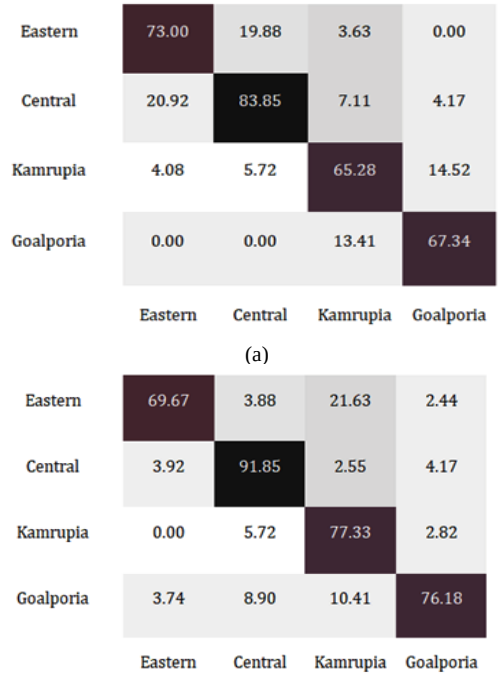


Fig. 10. Confusion matrix with (a) dynamic features (b) static features.

## V. CONCLUSION

The study focused on the analysis and categorization of vowels in four Assamese dialects. It examined the acoustic and prosodic characteristics of eight vowels, both individually and when combined together. The research involved extracting acoustic characteristics, specifically formant frequencies F1, F2, and F3. Additionally, longer frame prosodic features were analyzed as well. The study revealed that combining formant frequencies (F1, F2, and F3) with pitch (F0), energy/loudness, and duration led to better results. Three ensemble algorithms based on decision trees, namely RF, ERF, and XGB, were utilized for classifying the

Assamese dialect using vowel data. Statistical analyses using ANOVA highlighted the significant contributions of formant frequencies (F1, F2, and F3) in the categorization of the Assamese dialect. The study extracted acoustic features from vowels to evaluate how they exhibited dynamic and static characteristics across different dialects. The performance of statistical features produced from statistical parameters was higher, and the accuracy and consistency of ERF were better than those of RF and XGB. Additionally, it was discovered that static features outperformed dynamic features produced by contour trends.

In order to classify vowel-based dialects more accurately, it was crucial to examine the specific contributions made by each characteristic. This study conducted an examination of the contributions of each feature, both individually and collectively. Cross-Validation (CV) results from the ERF ensemble model were given in Tables IX–XI. When taking into account the features, the ERF method outperformed the other two algorithms (as shown in Fig. 9). Several interesting observations were made: The roles of the three formants were observed to be comparable, thereby confirming the findings obtained from the Analysis of Variance (ANOVA).

TABLE IX. THE FUNCTION OF FORMANT CHARACTERISTICS IN RECOGNIZING DIALECTS

| Features | Formants (63.44%) | | |
|---|---|---|---|
| | F3(1) | F2(2) | F1(3) |
| Total Contribution (%) | 34.97 | 34.27 | 34.20 |

TABLE X. THE FUNCTION OF THE PROSODIC FEATURES IN RECOGNIZING DIALECTS

| Features | Formants (63.44%) | | |
|---|---|---|---|
| | Duration(1) | Energy(2) | Pitch(3) |
| Total Contribution (%) | 37.52 | 34.97 | 30.69 |

TABLE XI. THE FUNCTION OF THE COMBINED FEATURES IN RECOGNIZING DIALECTS

| Features | Formants (63.44%) | | | | | |
|---|---|---|---|---|---|---|
| | Duration(1) | Energy(2) | Pitch(3) | F3(4) | F2(5) | F1(6) |
| Total Contribution (%) | 33.23 | 31.97 | 16.74 | 15.91 | 13.84 | 13.70 |

With regard to prosodic characteristics, duration was found to be the most influential, followed by energy and F0, highlighting the significance of differences in speaking duration among dialects. F0 made a relatively smaller contribution in comparison. When the features were combined, a pattern emerged where duration, energy, pitch (F0), formants (F3, F2, and F1) demonstrated the most significant contribution to dialect classification using Assamese vowels.

The present study aimed to investigate the relationship between the classification of Assamese dialects and the acoustic and prosodic characteristics of vowels. The classification of four Assamese dialects depended on the production of vowels in both read and spontaneous speech. Acoustic features were extracted to analyze the dynamic and static acoustic behavior of vowels, and ensemble algorithms were employed to achieve more accuracy in recognizing dialects. The findings revealed that although vowels exhibited varying formant frequencies, these alone were insufficient to distinguish between Assamese dialects. However, through statistical analysis, it was determined that the duration feature of vowels, showing the vowel's rate of speech, made a significant contribution among the considered acoustic features (as observed in Tables IX–XI). Notably, the Central dialect demonstrated better categorization with the inclusion of both dynamic and static features. These results highlighted the significance of the duration attribute in classifying different dialects. Among the remaining prosodic features, intensity contributed more significantly than the F0 feature. In general, the integration of prosodic characteristics and formants resulted in enhanced recognition of dialects based on vowel sounds.

A significant finding of this study is the potential of using formant and prosodic features to differentiate between dialects. This research provides the basis for future efforts to develop an effective method for identifying dialects and languages. Such systems have the potential to be designed for consumer-level or edge devices, facilitating localization of content and language-based service selection.

Future research directions can prioritize the exploration of specific distinguishing characteristics of dialects, that is, speech intonation, pace, and rhythmic patterns. It would be helpful to accurately analyze vowel onset and offset locations in the Assamese dialect to further enhance performance. Additionally, by carefully choosing optimized hyper parameters for ensemble classifier techniques, dialect identification performance can be improved. Moreover, a valuable endeavor would involve identifying characteristics of dialects that are not solely reliant on the underlying language, thus expanding the understanding of dialect variation.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

Hem Chandra Das conducted the complete research, formulated the initial manuscript, planned the study, conducted a comprehensive literature review, and played a role in analysing and interpreting the data. Utpal Bhattacharjee offered valuable insights for refining the manuscript, participated in analysing and interpreting the data, aided in writing and creating figures, and provided assistance with the experiments; all authors had approved the final version.

## REFERENCES


[1] J. K. Chambers and P. Trudgill, *Dialectology*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 1998.

[2] E. Ferragne and F. Pellegrino, "Automatic dialect identification: A study of British English," *Speaker Classification II: Selected Projects*, vol. 4441, pp. 243–257, 2007.

[3] M. Najafian, A. DeMarco, S. Cox *et al*., "Unsupervised model selection for recognition of regional accented speech," in *Proc. 15th Annu. Allerton Conf. of the International Speech Communication Association*, Singapore, 2014, pp. 2967–2971.

[4] M. J. Harris, S. T. Gries, and V. G. Miglio, "Prosody and its application to forensic linguistics," *Linguistic Evidence in Security, law and intelligence*, vol. 2, no. 2, pp. 11–29, 2014.

[5] H. Li, B, Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[6] S. Gray and J. H. L. Hansen, "An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Mexico, 2005, pp. 35–40.

[7] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[8] M. Mehrabani and J. H. L. Hansen, "Automatic analysis of dialect/language sets," *International Journal of Speech Technology*, vol. 18, pp. 277–286, 2015.

[9] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Graduate School of Arts and Science, Columbia Univ., New Work, 2011.

[10] G. A. Liu and J. H. L. Hansen, "A systematic strategy for robust automatic dialect identification," in *Proc. 19th European Signal Processing Conf*., Spain, 2011, pp. 2138–2141.

[11] G. C. Goswami, *Structure of Assamese*, 1st ed. Dept. of Publication, Gauhati University, India. 1982.

[12] B. Bharali, *Kamrupi Upabhasha: Eti Adhyayan*, Banlata, Guwahati, Assam, India, 2008.

[13] U. Goswami, *A Study on Kamrupi: A Dialect of Assamese*, Dept. of Historical Antiquarian Studies, Assam, India, 1970.

[14] Resource Centre for Indian Language Technology Solutions, Indian Institute of Technology, Guwahati. [Online]. Available: https://egovindia.wordpress.com/2006/06/21/resource-centre-for-indian-language-technology-solutions-rcilts-iit-guwahati/assamese-language

[15] B. Bharali and K. Talukdar, *Goalpariya Upabhasha: Rup Boichitrya*, Kumarpara, Shib Prakashan, Guwahati, Assam, India, 2012.

[16] M. Sarma and K. K. Sarma, *Phoneme-Based Speech Segmentation Using Hybrid Soft Computing Framework*, New Delhi: Springer, pp. 77–92, 2014.

[17] G. C. Goswami and J. P. Tamuli, *Asamiya. The Indo-Aryan Languages*, London: Routledge, pp. 391–443, 2003.

[18] N. F. Chen, S. W. Tam, W. Shen *et al*., "Characterizing phonetic transformations and acoustic differences across English dialects," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 110–124, Jan. 2014.

[19] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using gaussian mixture models," in *Proc. ODYSSEY04, The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, vol. 2.

[20] T. Chen, C. Huang, and E. Chang, "Automatic accent identification using Gaussian mixture models," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'01*, Italy, 2001, pp. 343–346.

[21] N. B. Chittaragi, A. Prakash, and S. G. Koolagudi, "Dialect identification using spectral and prosodic features on single and ensemble classifiers," *Arabian Journal for Science and Engineering*, vol. 43, no. 8, pp. 4289–4302, Oct. 2017.

[22] J. L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1904–1911, Aug. 2007.

[23] F. Biadsy and J. B. Hirschberg, "Using Prosody and Phonotactics in Arabic Dialect Identification," *Interspeech*, vol. 9, pp. 208–211, 2009.

[24] K. S. Rao and S. G. Koolagudi, "Identification of Hindi dialects and emotions using spectral and prosodic features of speech," *International Journal of Systemics, Cybernetics and Informatics*, vol. 9, no. 4, pp. 24–33, 2011.

[25] N. Dehak, P. A. T. Carrasquillo, D. Reynolds *et al*., "Language recognition via i-vectors and dimensionality reduction," in *Proc. the 12th Annu. Conf. International Speech Communication Association*, 2011, pp. 857–860.

[26] J. H. L. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of accent and language information," *Speech Communication*, vol. 78, pp. 19–33, April 2016.

[27] H. Behravan, V. Hautamäki, and T. Kinnunen, "Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish," *Speech Communication*, vol. 66, pp. 118–129, Feb. 2015.

[28] Y. Lei and J. H. L. Hansen, "Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 85–96, Jan. 2010.

[29] M. A. Zissman, T. P. Gleason, and D. M. Rekart, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech," in *Proc. 1996 IEEE International Conf. on Acoustics, Speech, and Signal Processing*, 1996, pp. 777–780.

[30] R. Huang, J. H. L. Hansen, and P. Angkititrakul, "Dialect/accent classification using unrestricted audio," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 454–464, Feb. 2007.

[31] M. Sarma and K. K. Sarma, "Dialect identification from Assamese speech using prosodic features and a neuro fuzzy classifier," in *Proc. 3rd International Conf. on Signal Processing and Integrated Networks (SPIN)*, India, 2016, pp. 127–132.

[32] W. Shen, N. Chen, and D. Reynolds, "Dialect recognition using adapted phonetic models," in *Proc. 19th Annu. Conference of the International Speech Communication Association*, Australia, 2008, pp. 763–766.

[33] T. Purnell, W. Idsardi, and J. Baugh, "Perceptual and phonetic experiments on American English dialect identification," *Journal of Language and Social Psychology*, vol. 18, no. 1, pp. 10–30, March 1999.

[34] F. Biadsy, J. B. Hirschberg, and N. Y. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proc. the Workshop on Computational Approaches to Semitic Languages*, 2009, pp. 53–61.

[35] N. F. Chen, W. Shen, and J. P. Campbell, "A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models," in *Proc. 2010 IEEE International Conf. on Acoustics, Speech and Signal Processing*, 2010, pp. 5014–5017.

[36] P. Escudero, P. Boersma, A. S. Rauber *et al*., "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1379–1393, Sept. 2009.

[37] Z. Ge, "Improved accent classification combining phonetic vowels with acoustic features," in *Proc. 8th International Congress on Image and Signal Processing (CISP)*, 2015, pp. 1204–1209.

[38] C. G. Clopper, D. B. Pisoni, and K. D. Jong, "Acoustic characteristics of the vowel systems of six regional varieties of American English," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1661–1676, Sept. 2005.

[39] P. Adank, R. V. Hout, and R. Smits, "An acoustic description of the vowels of Northern and Southern Standard Dutch," *The Journal of the Acoustical society of America*, vol. 116, no. 3, pp. 1729–1738, Sept. 2004.

[40] C. Themistocleous, "Dialect classification using vowel acoustic parameters," *Speech Communication*, vol. 92, no. 3, pp. 13–22, Sept. 2017.

[41] A. B. Ximenes, J. A. Shaw, and C. Carignan, "Dialect classification using vowel acoustic parameters," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 363–377, July 2017.

[42] M. Sarma and K. K. Sarma, "Dialect identification from assamese speech using prosodic features and a neuro fuzzy classifier," in *Proc. 3rd International Conf. on Signal Processing and Integrated Networks (SPIN)*, India, 2016, pp. 127–132.
[43] S. G. Koolagudi, D. Rastogi, and K. S. Rao, "Identification of language using Mel-Frequency Cepstral Coefficients (MFCC)," *Procedia Engineering*, vol. 38, pp. 3391–3398, 2012.
[44] V. K. Verma and N. Khanna, "Indian language identification using k-means clustering and support vector machine (SVM)," in Proc. *2013 Students Conf. on Engineering and Systems (SCES)*, 2013.
[45] T. Ismail and L. J. Singh, "Dialect identification of assamese language using spectral features," *Indian Journal of Science and Technology*, vol. 10, no. 20, pp. 1–7, May 2017.
[46] S. Sinha, A. Jain, and S. S. Agrawal, "Fusion of multi-stream speech features for dialect classification*," CSI Transactions on ICT*, vol. 2, pp. 243–252, June 2015.
[47] I. T. Utami, B. Sartono, and K. Sadik, "Comparison of single and ensemble classifiers of support vector machine and classification tree," *Journal of Mathematical Sciences and Applications*, vol. 2, no. 2, pp. 17–20, 2014.
[48] R. Chitturi and J. H. L. Hansen, "Multi-stream dialect classification using SVM-GMM hybrid classifiers," in *Proc. 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 431–436.
[49] N. E. Lachachi and A. Adla, "Two approaches-based L2-SVMs reduced to MEB problems for dialect identification," *International Journal of Computational Vision and Robotics*, vol. 6, no. 1–2, pp. 1–18, Dec. 2016.
[50] N. B. Chittaragi and S. G. Koolagudi, "Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms," *Language Resources and Evaluation*, vol. 54, pp. 553–585, 2020.
[51] H. C. Das and U. Bhattacharjee, "Identification of four major dialects of Assamese language using GMM with UBM," in *Proc. 3rd International Conf. on Machine Intelligence and Signal Processing*, India, 2021, pp. 311–319.
[52] P. Sarmah and L. Dihingia, "Assamese dialect identification from vowel acoustics," in *Proc. 2021 Conf. of Data Engineering for Smart Systems*, 2022, pp. 313–322.
[53] N. B. Chittaragi, A, Limaye, N. T. Chandana *et al*., "Automatic text-independent Kannada dialect identification system, in information systems design and intelligent applications," in *Proc. 5th International Conf. Information Systems Design and Intelligent Applications*, India, 2019, pp. 79–87.
[54] F. Biadsy, J. Hirschberg, and P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proc. 12th Annual Conference of the International Speech Communication Association*, 2011, vol. 3.
[55] K. Darwish, H. Sajjad, and H. Mubarak, "Verifiably effective Arabic dialect identification," in *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1465–1468.
[56] S. Malmasi and M. Dras, "Language identification using classifier ensembles," in *Proc. the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, 2015, pp. 35–43.
[57] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341–345, 2001.
[58] H. Reetz and A. Jongman, *Phonetics: Transcription, Production, Acoustics, and Perception*, 2nd ed. New York: Wiley, 2020, pp. 206–228.
[59] Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Hall Englewood Clifs: PTR Prentice, 1993.
[60] S. S. Agrawal, A. Jain, and S. Sinha, "Analysis and modeling of acoustic information for automatic dialect classification," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 593–609, July 2016.
[61] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," *The Journal of the Acoustical Society of America*, vol. 105, no. 1, pp. 512–521, Jan. 1999.
[62] V. R. Reddy, S. Maity, and K. S. Rao, "Identification of Indian languages using multi-level spectral and prosodic features," *International Journal of Speech Technology*, vol. 16, no. 1, pp. 489–511, May 2013.
[63] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," in *Proc. 6th International Conference on Spoken Language Processing*, 2000, vol. 5.
[64] R. K. Aggarwal and M. Dave, "Using Gaussian mixtures for Hindi speech recognition system," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 4, no. 4, pp. 157–170, Dec. 2011.
[65] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. International Workshop on Multiple Classifier Systems*, 2000, pp. 1–15.
[66] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[67] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd International Conference on Knowledge Discovery and Data Mining*, USA, 2016, pp. 785–794.