Within-formant spectral feature analysis for forensic speaker discrimination casework:

A study of 45 Marwari monolinguals from Bikaner, India

Nikita Suthar

PhD

University of York

Language and Linguistic Science

October 2023

Abstract

This PhD project investigates the significance of within-formant measurements for the vowels [iː], [ɪ], [e], [ə], [aː], [o], [uː], and [ʊ], for forensic speaker comparison. It contains six traditional PhD thesis chapters providing background information, as well as three research articles presenting analyses.

Data was sourced from the Marwari language, spoken in Rajasthan, India, as a testbed, but its applicability may extend to other languages. Speech was recorded from forty-five female Marwari monolingual speakers representing three caste dialects (fifteen per variety). Three speech elicitation techniques were used: reading from a wordlist, telling stories around picture stimuli, and engaging in conversation.

Articles 1–3 investigate the impact of including within-formant spectral moments (i.e., centre of gravity, standard deviation, kurtosis, skewness) and spectral measures (i.e., formant amplitude, relative amplitude, spectral bandwidth, LPC bandwidth, and spectral peaks), with and without centre formant frequencies, on speaker discrimination models. The investigations encompass various combination-based systems tested against three separate variables - vowels, variety, and speech style - using linear mixed model ANOVA and linear discriminant analysis.

The research contributes to existing manual systems by providing a semi-supervised feature-based system that may supplement existing ‘manual’ and semi-supervised tools. For legal systems that currently do not accept ASR analysis, it provides a more interpretable and reproducible approach.

Table of Content

List of Figures

List of Tables

Dedicated to my Nanaji and Papa…...

You have been my constant inspiration. I only wish you could see me now…

Acknowledgments

It is very “crucial” that first and foremost I acknowledge and sincerely thank my supervisor, Professor Peter French for his invaluable guidance and feedback at every stage of my research. Our thought-provoking discussions never failed to motivate me to enrich my thinking and strengthen my efforts. I could not have been able to pursue this PhD, let alone produce this thesis without his patient support and academic rigour. I am extremely grateful for his kindness and understanding when I made mistakes, as well as his insights in guiding me to correct these mistakes. The last four years have been filled with ups and downs, and I am sincerely grateful to Professor French for being a constant pillar of encouragement and support along the way. I also wish to thank Professor Dominic Watt for generously joining our supervision team during the final, critical year of this project.

I am also extremely thankful to the Wolfson Foundation for providing the funding that enabled me to fully focus on my PhD research over the past few years. This work would not have been possible without their generous assistance.

My heartfelt gratitude goes to Dr Philip Harrison for helping me with the most difficult task I had, namely Praat scripting. I could not have finished my project on time without his help. I would also want to thank Dr Vincent Hughes for graciously offering helpful comments as my Thesis Advisory Member and for helping and supporting me with his statistical knowledge whenever I needed it-you made my work so much stronger.I would especially like to thank Dr Justin Lo and Dr Bruce Wang for their patience and willingness to address my queries related to LLR on very short notice.

I would also like to thank Dr Claire Childs and Dr Rhys Sandow, for giving me the opportunity to work as a GTA. Seminar classes were my sweet escape from all the PhD hassle, and I enjoyed every bit of it (including marking).

I would also like to thank my informants and participants for helping me navigate these close-knit communities and welcoming me with open arms.

Completing a PhD during a global pandemic, while physically disconnected from everyone you care about, has been the greatest challenge of my academic journey so far. However, the support I received from all of you kept me going through the difficult times and helped me successfully finish this milestone.

I feel really lucky to have had Stuti and Alison by my side throughout this whole journey. Our walks and those spontaneous taco nights/BBQs were like a breath of fresh air when things got hectic. Alison's sharp eye on my drafts? That was a game-changer, seriously. I can't even begin to express how much I appreciate their unwavering friendship, whether I was on top of the world or feeling overwhelmed.

The PhD journey can be long and lonely but going through it with your best friend is not only bearable but also enjoyable. Stuti Bhagat, a huge thank you for being not only the kindest soul with a South Delhi accent but also the second-most amazing cook (we all know who takes the top spot). The warmth and culinary talents that you brought with you to the UK enriched this journey in countless ways. I want to give a big shoutout to my friend Lukas, who has always been there for me, even with his busy schedule (and providing me valuable feedback on how to navigate life.)

I would like to extend a special thank you to '*my sister from another mister*' Dr Maria Gabriela for keeping me grounded and preventing me from turning into a nervous wreck. You're a lifesaver!

I would also like to thank and acknowledge my awesome FSS colleague Samantha, thanks for being the patient ear to all my endless analysis questions (and shopping-related questions) and giving me your expert advice.

I am deeply grateful to my mother and Prince for their unwavering support and encouragement from India. Words cannot express my appreciation for your unconditional love. Thank you for your incredible patience in dealing with my stress-induced tantrums and always being there to calm me down. It's hard to believe that it's been twenty years since I left home, but both of you have made sure that I never felt homesick, no matter where life took me. Mom, your belief in the limitless possibilities of life has been a constant source of inspiration. And Prince, your annoying habit of reminding me of impending deadlines, has kept me on track. Thank you both for being such a vital part of my life.

Sascha, having you around in York made it feel like home sweet home. Our adventures across Europe taught me so much (and the learning journey is far from over). You are the most patient person I have ever met. Thank you for giving me a crash course on how to be patient while

also being stressed at the same time. Thank you so much for teaching me how to make the best coffee (and the significance of coffee).

I couldn’t have made it through this without such an amazing support system. And to all my dear friends in India and Europe, you’ve kept me smiling through these tough years. My heartfelt thanks to each one of you who played a part in this achievement. I’ll always remember and appreciate your support.

With appreciation

Nikita

Declaration

This thesis has not previously been submitted for any degree other than Doctor of Philosophy of the University of York. This thesis is only my original work, except where otherwise stated. Other sources are acknowledged by explicit references. This is a hybrid-style thesis where three articles have been included as separate chapters. The following elements of the study has been submitted for publication as articles and are under review.

- Chapter 6 presented as Article 1:

  Suthar, N. & French, P. (under review) Primary spectral moments of the first four vowel formants as a source of speaker discriminant information. *Speech Communication.* https://dx.doi.org/10.2139/ssrn.4581148

- Chapter 7 presented as Article 2:

  Suthar, N. & French, P. (under review) Within-formant spectral measures and their role as a source of speaker discriminant information. *The International Journal of Speech, Language and the Law.*

1 Introduction

This work, submitted in accordance with the guidelines for the degree of PhD at the University of York, has two parts. The first part comprises five chapters written in the format of a traditional/standard United Kingdom PhD thesis. The second part consists of three publishable journal articles that report on the results of the research. Two of these articles (Articles 1 and 2) have already been submitted for peer-review to two different journals. The final chapter of the thesis (chapter 9) provides an overall conclusion of the study (which is written in the form of a traditional/standard UK PhD thesis). Because the research consists of standard thesis-style chapters and articles, the term "study" will be used instead of "thesis."

I began the research with the aim of advancing the acoustic component of auditory-cum acoustic-phonetic speaker comparison by investigating the value of adding within-formant spectral measures (formant amplitude, formant bandwidth and spectral peak) and spectral moments (centre of gravity, standard deviation, kurtosis and skewness) to the existing test battery.

Given the time-consuming and labour-intensive nature of extracting these features, one might question why not simply leverage the sophisticated automated methods offered by various modern ASR software. There are three key rationales examining these measures, which serve as responses to the aforementioned question:

Reason 1: If one accepts the proposition that formants can carry information about individual speakers' vocal tract geometry, which at least some of the earlier work is based on (Cavalcanti et al., 2021; Nolan & Grigoras, 2005; Traunmüller, 1984), then there may be a good reason to think that speaker individuality may manifest in subtler, more nuanced acoustic dimensions of the formants than just centre frequencies, namely spectral measures and spectral moments. There is relatively limited work available for these features and the present study hypotheses that by examining additional minute characteristics of this already proven measure, it is possible to shed some light on the facets of formants that are actually responsible for speaker characterisation, and whether focusing on these elements in detail, along with formants, can improve the system's efficiency.

Reason 2: ASR systems extract both spectral features (the frequency contents of the speech signal) and temporal features (change in speech over time) from a speech signal that is relevant

to the task of speech recognition. However, these choices are difficult to interpret, as these systems are typically trained on large datasets of speech data, and the decision-making process is often hidden within the neural network that is used to train the system. The present study's focus is on the minute spectral features, which are explainable and can be replicable, thus providing grounds for a more interpretable system, which is based on the features that have been proven to present inter-speaker differences.

Reason 3: The admissibility of ASR-derived results as evidence is a complex issue. There are countries where the evidence produced by an ASR system is accepted, for example, the USA, New Zealand, Australia and Canada. However, there are many countries that do not prefer or completely reject any evidence produced by an ASR system. For example, evidence produced by automatic speaker recognition (ASR) systems was not admitted by the England & Wales Court of Criminal Appeal in the case of R v Slade & Ors (2015). Since England and Wales are a Common Law jurisdiction, this ruling at least makes it very difficult to have ASR evidence accepted by a lower court at present. Moreover, the admissibility of ASR-derived results as evidence is still an issue of debate for Indian courts, and many other Commonwealth countries, and the law in this area is still developing. In the absence of a local ruling, courts faced with ASR evidence in any of the other 57 Commonwealth countries may look to the England and Wales ruling for guidance (French, 2017). Unless and until the legal position changes, the development and strengthening of the auditory-cum acoustic-phonetic method of forensic speaker comparison must continue. Considering all these factors, having a more interpretable system can be more beneficial in such cases.

Within-formant measurements are time-consuming if carried out using the methods I used here. I embarked on the research with the full knowledge that this would be the case. Of course, one could argue that the measurements are so time-consuming that they would never be used by forensic speech scientists in their casework. All the within-formant measures can be automated, but that would be a second step–a step that would only be worth taking once the measurements are demonstrated to have speaker-discriminatory value. This is what the present research seeks to establish.

The research reported here is intended to contribute to that process.

The reader may ask what motivates research into within-formant measurements. Why should anyone think that these particular measurements hold speaker-discriminatory value? The most

honest and straightforward answer is that we do not know. It is not currently possible to relate energy kurtosis or negative skew to particular vocal tract settings or articulatory orientations. One can question if it is appropriate to term these features explainable if they cannot be explained by a person's vocal tract geometry, as shown in Reason 2 (cf. above). One answer to this could be that although these measurements cannot yet explain specific vocal tract configurations, they are retrieved and computed from a speaker's voice spectrum and indicate the acoustic energy inside their formants. The relationship to vocal tract could be addressed via different kinds of empirical study at a later point once the discriminatory value has been established. There is a history of this in forensic phonetics and acoustics, the major example being the use of Mel Frequency Cepstral Coefficients (MFCCs), the units that form the basis for most ASR systems. MFCCs were not designed to discriminate individual speakers, but to assist with speech recognition (Bhatt et al., 2021; Davis & Mermelstein, 1980). It was at a later point that they were discovered to have individual discriminatory values (Campbell, 1997; Gish & Schmidt, 1994). I still have seen no convincing explanation of how they relate to vocal tract dimensions or configuration, such that one could say, for instance, that the seventh coefficient is related to tongue body fronting or that the fourth relates to the size of the nasopharyngeal cavity. Despite this, their value has been repeatedly demonstrated by experimental studies and trials now showing blind testing equal error rates of less than one percent (Das & Li, 2020; Kajarekar et al., 2009).

I return to these questions in the conclusion chapter at the end of the study. The following section provides a summary overview of the chapters and the articles.

1.1 Overview of the Study

1.1.1 Chapter 1: Introduction

Chapter 1 starts by providing a brief rationale for the study. It also summarises each chapter and article, while also providing an outline of the structure of the study.

1.1.2 Chapter 2: Background

Chapter 2 is an introduction to the field of forensic speaker comparisons as well as to the techniques and methodologies employed by forensic phoneticians. This chapter also summarises the various within-formant features and their analysis methodologies used for the

present investigation. The chapter outlines the role and background of formant-based analysis as a source of speaker discrimination. It also presents an overview of automatic speaker recognition systems. Finally, it discusses the rationale and purpose of the current study and how it complements the established methodologies.

1.1.3 Chapter 3: Marwari language

Chapter 3 provides a general introduction to the Marwari language, including a synopsis of its phonetics, phonology and morphosyntactic structure. The study includes speech samples representing three different caste-dialects of Marwari. As background, an overview of the three separate Hindu castes and their origins is provided in the chapter. Based on an analysis of the fieldwork recordings made for the present study, the final section of the chapter draws distinctions between each of these caste variants.

1.1.4 Chapter 4: Data collection

Chapter 4 discusses the methods employed during fieldwork. It also explains the rationale for the different kinds of data collected for the study. The chapter begins by outlining the participants' background and then moves on to discuss the fieldwork materials and ethical clearances.

1.1.5 Chapter 5: Data processing

Chapter 5 provides an overview of one of the main preparatory works undertaken to enable analysis of the speech data. This included isolating target sounds and extracting within-formant features. This chapter covers the justification for each preparatory decision and explains the methodology used.

1.1.6 Chapter 6/ Article 1: Primary spectral moments of the first four vowel formants as a source of speaker discriminant information

Chapter 6/ Article 1 is formatted in the style of the journal to which it was submitted: "Speech Communication (Online ISSN: 1872-7182)[1]."

[1] web address: https://www.sciencedirect.com/journal/speech-communication

It presents the results of within-formant analysis of spectral moments for eight different vowels. The spectral moments are examined individually and in combination. Additionally, the study investigates three variables: vowels, varieties and speech styles, in an effort to assess which of these provides better results for speaker discrimination work.

1.1.7 Chapter 7/ Article 2: Within-formant spectral measures and their role as a source of speaker discriminant information

Chapter 7 /Article 2 is formatted in the style of the journal to which it was submitted: "The International Journal of Speech, Language and Law (Online ISSN: 1748-8893)2."

Article 2 evaluates the role of within-formant spectral measures, including amplitude, bandwidth, and spectral peaks as a source of speaker discriminant. The article also discusses how these measurements, individually and in combination, affect speaker discrimination. Further, it shows which vowel, variety, or method of data elicitation yields the best results.

1.1.8 Chapter 8/ Article 3: Enhancing forensic speaker discrimination: a comprehensive spectral feature analysis of Marwari vowels using within-formant measures and spectral moments.

Chapter 8 /Article 3 is formatted in the style of the journal to which it will be submitted: "Speech Communication (Online ISSN: 1872-7182)[3]."

It integrates spectral moments and spectral measures in a new model. The impact of this model on the speaker classification outcomes is discussed. Additionally, the results of the new model are tested against the three variables, vowel, variety, and speech style.

1.1.9 Conclusion

Chapter 9 begins with a brief synopsis of each article. The second section of the chapter contains the study's general summary and conclusion. The chapter finishes with a summary of the study's general limitations and implications.

[2] Web address: https://journal.equinoxpub.com/IJSLL

[3] Web address: https://www.sciencedirect.com/journal/speech-communication

2 Background

This chapter provides a brief introduction to acoustic phonetics, forensic speaker comparison and various types of systems employed in forensic speaker comparison cases, with a specific emphasis on those utilised in the present study. The chapter concludes by presenting an overview of the research questions posed in each article within this study.

2.1 Acoustic Phonetics

Acoustics is the discipline concerned with the physics of sound. Acoustic phonetics is the subset of the discipline that is concerned with physically describing a sound or entire speech signal and trying to explain the characteristics that account for its linguistic and auditory representation (Martin, 2021, p. 2). One of the theories to describe human speech production is the source-filter theory (Fant, 1971). The mechanism of speech production is characterised as a two-stage process in the source-filter theory: (a) The airflow from the lungs leads to tissue vibrations in the vocal folds and generates the "source" sound. (b) The vocal tract "filter" shapes the spectral patterns of these source sounds. Section 2.3.1 will discuss source-filter theory in further details.

Prior to establishing the various components of source-filter theory, a concise introduction to the domain of forensic speaker comparison is presented. This introduction serves the purpose of elucidating the rationale behind the utilisation of various acoustic measurements in the present study, as well as illustrating the interconnectedness of measuring and utilising these attributes for speaker discrimination within the broader context of forensic speaker comparison.

The subsequent section briefly outlines the field of forensic speaker comparison.

2.2 Forensic Speaker Comparison

Forensic speech science deals with various speaker and speech-related cases with the help of several approaches, including speaker profiling, speaker recognition, and speaker discrimination. Forensic speaker or voice comparison is one of the subfields of forensic speech science. Phoneticians or speech engineers frequently conduct forensic speaker or voice comparison tasks to assess the likelihood that two or more speech recordings originate from the same individual (Foulkes & French, 2012). The experts use their knowledge in phonetics, acoustics, signal processing or statistics to select from a variety of methods to carry out any

speaker or voice comparison tasks in order to support or refute the evidence for legal proceedings (Jessen, 2008; Rose, 2002). This kind of analysis includes assessing a retrieved voice recording of an unknown offender during the crime (question/criminal sample) against the reference samples acquired from a suspect (known/suspect sample). Although suspect samples are typically acquired during police interviews, they can also be retrieved separately for speaker comparison tasks. The various suspect samples rely on the legal requirements of a given legal system, which can differ from country to country. In certain countries, for example in the UK, it is a legal requirement to record police interviews (Home Office, 2017). In other countries, for example in India, police interview records are not admissible as evidence, and it is not mandatory for them to be recorded. In such cases a known sample from the suspect is collected later by the police/criminal labs. These circumstances involve the acquisition of suspect recordings as part of the evidence collected for a specific purpose, which may have an impact on the recordings since the suspect may alter or modify their speech samples (Alison et al., 2008). To account for such scenarios, more sophisticated and reliable techniques can be utilised to extract speaker-specific information irrespective of any external influences.

Here, it is important to note that the term "speaker comparison" rather than "speaker identification" is chosen for two reasons. Firstly, both manual and automatic approaches aim to offer information about the speech samples and speaker(s) rather than a speaker's identity (Foulkes & French, 2012). Secondly, by offering an identification, the expert should not assume the position of the trier of fact. Instead, it is his/her duty to indicate the likelihood of collecting the evidence under the premise that the samples were created by the same individual against the likelihood of acquiring the evidence under the assumption that the criminal and suspect samples were created by two distinct individuals. Consequently, the term "speaker" is preferred over "voice" since not all factors addressed in forensic speaker comparison studies are solely the results of the voice. As noted by French et al. (2010), some aspects of speech have more to do with language and non-linguistic behaviours than voice, which in turn, may be influenced by an individual's social and psychological attitudes (Wang, 2021). Based on this understanding, the analysis in a forensic speaker comparison has been divided into five different approaches by Gold and French (2011). A summary of the approaches is provided below:

Auditory Phonetic Analysis Only (AuPA): AuPA is a comparison based on expert listeners' auditory analyses of the segmental and suprasegmental aspects of the speech sounds.

Acoustic Phonetic Analysis Only (AcPA): AcPA is an analysis conducted by an expert based on the physical parameters of speech, which can be analysed and quantified with the help of computer program such as Praat (Boersma & Weenink, 2001). It is a labour-intensive and time-consuming approach, which differs from the automatic approaches in that it depends on expert supervision at every stage.

Auditory Phonetic cum Acoustic Phonetic Analysis (AuPA+AcPA): This approach combines AuPA and AcPA.

Automatic Speaker Recognition System (ASR): This method evaluates the degree of similarity between speech samples using statistical models of variables collected automatically from the recording with the help of a specialised software.

Automatic Speaker Recognition System with Human Assistance (HASR): This method evaluates the similarities or/and differences between two recordings with the help of an automatic speaker recognition system. This approach has undergone intensive training and calibration by specialists based on various feature extraction models and speech samples. With the aid of this pre-trained system's automated analysis, the degree of similarity is calculated for any speech sample. Human assistance can range from listening to the speech samples to custom training these models.

Morrison et al. (2016) adapted Gold and French's methods by further separating the HASR into two new categories, one where the automation is performed and verified by a forensic practitioner and one where it is performed and checked by non-practitioners such as police officers. They also featured two acoustic-based techniques, one of which relied on qualitative judgement through visual scans of spectrograms and the other on quantitative analysis. Table 2.1 provides a brief overview of these approaches along with the regions they are used in.

Gold and French (2011) and Morrison et al. (2016) acknowledge that there is no universal consensus on which methodology should be utilised or how these results should be presented, because the distributions of various methodologies fluctuate for different countries, regions and even for different types of workplaces. According to Gold and French (2011), the most often employed approach is AuPA+AcPA in the UK. A second survey revealed a rising tendency in the use of some form of HASR over time, however, AuPA+AcPA was used by majority of countries.(Gold & French, 2019). Morrison et al. (2016)'s survey, on the other

hand, found that both HASR systems and auditory-acoustic phonetic methods were used by every region to some extent.

This study utilises Acoustic-Phonetic Analysis (AcPA) (based on statistical modelling) which combines the AcPA approach proposed by Gold and French (2011) and acoustic-phonetic by forensic practitioners (statistical) proposed by Morrison et al. (2016). Articles 1–3 discuss the importance of exploiting various within-formant acoustic properties and propose numerous statistical models based on various parameters. Incorporating these features into the AcPA based on statistical modelling can improve the current battery of speaker comparison methodologies.

The overarching goal is to contribute a new semi-supervised spectral characteristics system that would supplement existing features. In addition to introducing novel feature combinations with the help of this system, this work will provide justifications for the adoption of this specific approach.

Table 2.1 Use of different approaches as provided by Morrison et al. (2016)

Approach	Methodology	Region
Auditory by forensic practitioners	Based on listening to speech recordings, phoneticians make qualitative opinions.	Europe, Africa, Middle East, South and Central America
Spectrographic or auditory-spectrographic by forensic practitioners,	Practitioners form qualitative judgments by visually examining recordings.	Europe, Asia, Africa, Middle East, South and Central America
Auditory-acoustic-phonetic by forensic practitioners (qualitative)	Phoneticians form qualitative judgements by auditory and acoustic analysis.	North America, Europe, Asia, Africa, Middle East, South and Central America
Acoustic-phonetic by forensic practitioners (statistical)	Phoneticians utilise quantitative data and statistical models to determine evidence strength.	North America, Europe, Asia, Africa, Middle East
Human-supervised automatic approaches by forensic practitioners	Phoneticians/signal-processing engineers meticulously select and prepare recordings, which are subsequently analysed using signal-processing techniques, including statistical models, to assess the strength of the evidence.	North America, Europe, Asia, Africa, Middle East, South and Central America
Fully automatic approaches by non-forensic practitioners	Police officers using completely automatic systems without significant training in essential areas such as phonetics, signal processing, quantitative modelling, and forensic evidence interpretation.	Europe, Asia, Africa, Middle East

The following sections provide a brief introduction to various features used for the study.

2.3 AcPA-Based Semi-Supervised System

The present study will delve into the content and methodology of the three articles, focusing on the extraction, refinement and analysis of acoustic features using Praat and R.

These features originated from the analysis of the first four formants, and throughout the thesis, they would be referred to as “within-formant features” to emphasize their origin and utility. Additionally, the term “manually extracted features” would be used to underscore the hands-on refinement process applied to these acoustic attributes. Furthermore, it is important to note that the term “semi-supervised feature” will be introduced and utilised in selected sections of this thesis, to signify that the features initially underwent extraction through a scripted process and were subsequently subject to manual corrections. This distinction in terminology helps elucidate the workflow and refinement stages applied to the acoustic features under examination.

The upcoming sections 2.3.1 - 2.3.3 will provide a brief introduction to source filter theory, formants and within-formant features. Additionally, it will also explore the diverse statistical analysis methods utilised for their thorough examination. The detailed extractions and refinement process will be extensively discussed in Chapter 5. Lastly, the analytical results will be provided in Articles 1-3.

The following section will discuss one of the foundational models utilised in speech signal analysis, i.e., the source-filter model.

2.3.1 Source-filter model

Fant's (1971) source-filter theory sheds light on how the vocal tract shapes speech and is central for understanding speech mechanics. It explains speech production by defining speech waves as being composed of the source (sound) and the filter (vocal cavities - pharynx, mouth and nasal passage) components. The glottal source travels through the vocal tract, acting as a filter to produce speech. The source sound is a complex waveform characterised by its fundamental frequency (f0) (Fant, 1971; Harrison, 2013), representing the time per pulse or cycle. It comprises harmonics, multiples of f0, which determine the source's spectral characteristics. Factors such as phonation and vocal effort can influence these characteristics (Harrison, 2013). The filter function or transfer function represents the frequency-dependent properties of the vocal tract's radiation (Fant, 1971). It essentially shapes the frequencies of the source, producing a filtered output.

Changes in the source, such as altering f0, impact the vibratory frequencies but do not affect the filter or resonator (Kent & Read, 2002), indicating that the source and filter operate independently. It should be emphasised, however, that the source-filter theory is predicated on the traditional view that the source and the filter are separate entities. There have been studies that indicate how the source, and the filter can interact with one another in specific scenarios (e.g., Flanagan and Landgraf, 1968; Zhang et al. 2006 (a,b)). These studies show that a source sound is impacted by the shape of the vocal tract as well as the acoustic feedback from the vocal tract. This type of source-filter interaction causes a variety of voice instabilities, such as rapid pitch jumps, subharmonics, resonance, quenching, and 'chaos' (Titze, 2007).

Although theoretical, source-filter independence is still the most widely-cited interpretation of the human speech mechanism; the source-filter hypothesis approximates normal human speech

well, with the source sounds only minimally impacted by the vocal tract filter. The model has been effectively used in speech analysis, synthesis, and processing (Atal & Schroeder, 1978). The ability to adjust the source (phonation) and the filter (articulation) independently is helpful for acoustic communications with language, which involves the representation of numerous phonemes with a flexible manoeuvring of the vocal tract configuration (Fitch, 2010).

The present research will primarily focus on the resonance frequency outputs of the filter, known as formants. Section 2.3 will discuss what formants are and explain their critical relevance to the current investigation.

2.3.2 Formants

The acoustic filter function of the supralaryngeal vocal tract allows maximum energy to pass through certain frequencies and suppresses the energy at certain frequencies. The frequencies at which the maximum energies can pass through and can be analysed as resonance peaks are called "formants" (Lieberman & Blumstein, 1988). The width of these peaks is known as bandwidths. These resonant frequencies are the result of the interaction between acoustic space (e.g., for human speech mechanism it will be the vocal tract) and the sound. This interaction results in a higher concentration of energy in these frequency regions, resulting in formants (Stevens, 2000, p. 132). The resulting speech still consists of f0 and their respective harmonics, but it is shaped by the filter function of the vocal tract and presents the resonance peaks.

Since the formulation of the source-filter theory of speech production, speech research has advanced from the anatomy of sound production to the identification of a human based on these sounds. Formants have always been the focal point of this research. They have helped analyse and synthesise vowels (e.g., Klatt, 1982; Peterson & Barney, 1952) and with vowel perception and synthesis studies (e.g., Bladon & Lindblom, 1981; Ito et al., 2001; Miller, 1984). Phoneticians and speech signal engineers also started analysing vowel formants in the context of the forensic speaker recognition or comparison framework (e.g., Cao & Dellwo, 2019; Fleischer et al., 2015; Gonzalez-Rodriguez, 2011; Kent & Vorperian, 2018; McDougall, 2006; McDougall & Nolan, 2007).

With vowels, the frequencies of the formants F1 and F2 determine which vowel we perceive and are responsible for the differences in quality among different vowel sounds. At any one

point in time (as with spectra) there may be any number of formants, but for speech, the most informative are the first four (2), appropriately referred to as F1, F2, F3, and F4.

The first formant (F1) in vowels is inversely related to vowel height: the higher the vowel, the lower the first formant (and vice versa) (Peterson & Barney, 1952). The second formant (F2) in vowels is inversely related to the degree of backness. The more front the vowel, the higher the second formant (although it is affected by lip-rounding) (Peterson & Barney, 1952). The distance between F1 and F2 is a better predictor of the degree of backness in vowels (Miller, 1989). The closer F1 and F2 are to each other, the more back a vowel is (Peterson & Barney, 1952).

Previous studies of speaker differentiation through vowel sounds have consistently identified higher formants as valuable sources of discriminative information. McDougall (2004) and Hughes (2013) have suggested that F3 exhibited the most pronounced contribution to speaker distinction, surpassing F1 and F2 in importance. This prominence has been attributed to F3's direct articulatory correlates, leading to systematic variations. For example, F3 has often been associated with lip rounding, which could result in decreased cross-sectional area and elongation at the vocal tract (West, 1999). Furthermore, the lowering of F3 has been associated with rhotic vowels (Ladefoged, 2006; Lindau, 1978), and articulator configuration rather than phonological classification (Alwan et al., 1997). F3 reduction has also been related to constriction in the oral and pharyngeal cavities (Delattre & Freeman, 1968), generating resonance in the front of palatal constriction (Drager & Hay, 2012). The significance of F3 also extends to voice quality and vocal settings (Biemans, 2000; Klatt & Klatt, 1990; Laver, 1994). The systematic variation of F3 can be the result of various lingual settings affecting the vocal tract architecture (Laver, 1994). Zheng et al. (2012) stated vowel F3 plays a significant role in determining minute accent-related differences based on their research on Liverpool and Birmingham accents.

While studies have focused on F3, F4 and higher formants have often been ignored or removed from vowel identification as they have been considered to carry more speaker-specific information instead. Speaker comparison studies, however, have started focusing on these higher formants to assess inter-speaker variations. One example of this is Cao and Dellwo's (2019) work on the significance of the first five formants for forensic speaker comparison. They indicated that for the data acquired with a good recording quality, the combination of F4

and F5 alone provided the highest speaker classification rates. Speaker classification studies, just like vowel classification studies, are limited to the first three formants. The reason for this was that most of the speaker classification and discrimination-related work came from telephone recordings and the bandpass limitation of these recordings was usually 3500 Hz, so it was only possible to analyse the first three formants (Cao & Dellwo, 2019). Recent technological advancements, especially the upward-extended frequency range associated with social media messaging, have made it possible to acquire data from higher frequencies. However, this is still limited to very specific types of recorded data where the cut-off frequencies are at or higher than 5500 Hz, which makes them still unusable for many forensic recordings obtained from mobile or telephone recordings.

In recent years, human-assisted acoustic phonetic-based speaker recognition systems have expanded their feature analysis for improved outcomes. These include voice quality (Braun et al., 2021), pitch (Smith, 2016), and formant-related measures (Becker et al., 2008; Burris et al., 2014; Byrne & Foulkes, 2004; Cavalcanti et al., 2021; Ekaterini et al., 2016; Harrington et al., 2007), such as long-term formant distributions (Gold et al., 2013; Hughes et al., 2019).

To actively contribute to the ongoing progress in feature analysis, the present study strategically centred its attention on within-formant features. The main reasons for selecting these features were presented in the introduction chapter and are summarised here again:

Reason 1: Earlier work has demonstrated that formants may contain information about individual speakers' vocal tract geometry, which can also indicate that speaker individuality may be manifested in subtler acoustic dimensions, such as spectral measures and spectral moments, beyond centre frequencies.

Reason 2: Some of the uninterpretable characteristics that ASR systems look at can be predicted by looking at more subtle within-formant based measures.

Reason 3: Using a manually extracted feature-based system along with already established parameters such as formant frequencies, f0, LTFs, can help improve the system accuracy.

In this regard, the subsequent section will provide a comprehensive exploration of two specific within-formant feature groups: spectral moments and spectral measures. These measures serve as pivotal components of the investigation approach, designed to enhance understanding and provide advancements in the field of AcPA-based forensic speaker comparison.

2.3.3 Within-formant features

The study incorporated two kinds of spectral features: spectral moments and spectral measures, which offer distinct methods to analyse speech traits. Spectral moments involve statistical measures revealing energy distribution across frequencies in a sound signal, aiding the understanding of frequency spectral utilisation ("Frequency spectral utilisation" refers to how effectively or efficiently different frequency components are utilised or distributed within a sound signal).

The second feature, termed spectral measures, encompasses amplitude, bandwidth and spectral peaks, characterising resonance frequencies by their strength (amplitude), frequency range width (bandwidth) and dominant frequency (spectral peak).

In summary, while spectral moments provide statistical insights into energy distributions across frequencies, spectral measures focus more on analysing the characteristics of specific frequency components, notably formants. Sections 2.3.2.1 and 2.3.2.1 will provide concise overviews of these features.

2.3.3.1 Spectral moments

The initial feature set consists of spectral moments, which represent numerical distributions of spectral energy and can be computed from multiple regions. This study examines the first four moments: the centre of gravity, variance, skewness and kurtosis (Jongman et al., 2000; Kardach et al., 2002).

Centre of gravity (spectral moment 1 or m_1): the mean or centre of gravity indicates central spectral energy of any given spectrum, in this case a formant.

Variance (spectral moment 2 or m_2): the dispersion of spectral energy.

Skewness (spectral moment 3 or m_3): the symmetry of the distribution, with positive skewness showing the spectral tilt towards a higher concentration of energy on the lower frequencies and vice versa.

Kurtosis (spectral moment 4 or m_4): the peakedness of the spectrum. A peaky or well-defined peak would have positive kurtosis and a flatter or more plateau-like 'peak' would have negative kurtosis.

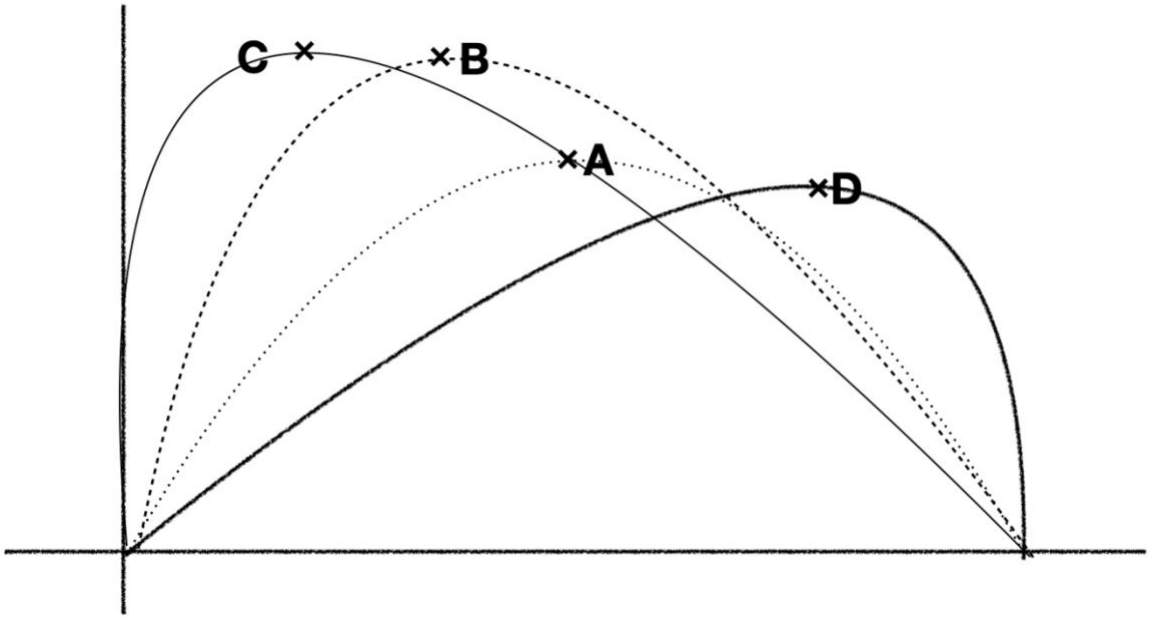


Figure 2.1 Four functions with varying degrees of skewness. The points represent the function's centre of gravity. Function A is symmetric denoting 0 skewness. Functions B and C are skewed to the right, resulting in positive skewness values, with function C being more skewed than B, denoting greater value. Function D is skewed toward the left demoting negative skewness.

These moments collectively depict the spectrum's overall shape. Künzel (2001) depicted limitations of formant-based speaker discrimination, and with the help of his work on telephone transmission showed that using formants alone is not as effective as we might think. In his study, he discussed two issues that affect the formant frequencies in a telephonic transmission. Firstly, the lower cut-off frequency of standard telephone transmission (around 400 Hz) attenuates lower frequency components, affecting the measurement of F1 for most vowels in adult male speech. Specifically, the higher frequency components get relatively more weight in determining F1, shifting its measured centre frequency upwards. Secondly, automatic formant extraction algorithms often fail for telephone speech. Reliable formant analysis requires manual verification, especially for forensic or dialectal studies. He summarised it as indicating that the telephone effect on F1 measurement is significant and can introduce errors in applications relying on precise formant data like speaker recognition and dialectology. His findings emphasised the limitations of using only formants for analysis of telephone speech. This problem was addressed by Rodman et al. (2002) through spectral moments. Their

approach combined centre formant frequencies and spectral moments for text-independent speaker comparison. This method adhered to conditions such as open-set assessments, handling short utterances, accommodating low signal-to-noise ratios, text independence, and relaxed time constraints (Rodman et al., 2002, p. 24). Kardach et al. (2002) extended this concept by defining the distribution of frequencies as spectral moments.

For the present study, these moments will be extracted from a spectral slice adjacent to either side of the peaks of the first four vowel formants (Further details will be provided in Chapter 5).

2.3.3.2 Spectral measures

The second feature set comprised amplitude, bandwidth, and spectral peaks.

Amplitude: the waveform deviation from the zero-line resulting from the air pressure changes during sound production.

Bandwidth: A spectral band indicating wideness or narrowness in the frequency range over which the formant energy is distributed. Speech sound bandwidth, or the span of frequencies occupied, is impacted by radiative properties as sound emanates from the lips as well as mechanical parameters along the entire vocal tract system including viscosity, heat conduction, constriction degree, and glottal state. Thus, many factors spanning respiration, phonation, resonance, and radiation contribute to the bandwidth shaping the frequency profiles of resulting speech sounds (Lindblom & Sundberg, 2014). A wider formant bandwidth suggests the energy is spread over a broader range of frequencies, while a narrower bandwidth indicates a more concentrated energy distribution within a specific frequency range. Further discussion on how bandwidth is calculated for the present study will be presented in Chapter 7/article 2.

Spectral peaks: the nearest maximum of estimated formants, i.e., the highest peak of the estimated formant. Spectral peaks are dependent on the level of cepstral smoothing applied on the spectrum. The extraction occurs after the LPC filter is applied to the speech signals (Rossing, 2014). The values of spectral peaks can coincide with formant peaks in some situations, but as the former are dependent on the cepstral smoothing and the filter, they are considered different from each other.

These features, termed “spectral measures,” are acoustic properties that provide important perceptual cues for vowel identification and discrimination. Amplitude is correlated with

formant frequencies, and any change in either of them can affect the quality of the vowels (Kent & Read, 2002). Speaker-specific information akin to formants might be linked to amplitudes. Kiefte et al. (2010) discusses that when formant peaks closer than 3-3.5 Bark scale coalesce into a single spectral prominence, additional spectral features such as the formant amplitude may alter perceived vowel qualities. Similarly, formant bandwidth variations convey speaker sex. Any substantial shift in formant bandwidth can affect the naturalness of the speech (Hawks & Miller, 1995; Kent & Read, 2002).

It should be emphasised that, while both formant bandwidth and amplitude variations can be perceptually relevant for the human ear, they are far less significant than formant frequencies. For example, Carlson et al. (1979) argued that the human ear was 20 times more responsive to changes caused by formant frequencies than formant bandwidths. They conducted perceptual studies to determine how sensitive human hearing is to changes in formant frequencies versus changes in formant bandwidths, in the context of vowel sounds. They found that listeners were much more sensitive to slight variations in formant frequencies compared to equivalently slight shifts in the bandwidths of those same formant resonances. Specifically, thresholds testing the just noticeable difference (JND) showed listeners could detect around a 3% shift in formant frequency. But formant bandwidths could vary by over 20% before listeners noticed the change. Based on this, they hypothesised that the auditory system seems over 20 times more sensitive to perturbations in the centre frequencies of vocal tract resonances, compared to equivalent relative changes in how narrowly or broadly the resonant energy is distributed around those centre frequencies.

Furthermore, there have been studies that support both sides of the argument, one claiming that spectral metrics such as formant bandwidth and formant amplitudes are perceptually important for vowels (e.g., Ainsworth & Millar, 1972; Carlson et al., 1979; Lindqvist-Gauffin & Pauli, 1968) and the other claiming that they are not (e.g., Assmann, 1991; Klatt, 1982). However, in both cases the focus of the investigation was on the significance of these measures for vowel perception rather than speaker discrimination.

While past research has revealed that neither amplitude nor bandwidth alone determine vowel perception, there is still a lack of consensus on the exact interplay and relative weighting of different acoustic cues. Global spectral properties beyond isolated formants likely contribute to identification and naturalness (Bladon & Lindblom, 1981; Ito et al., 2001; Miller 1984). At

the same time, the role of amplitude across speakers requires further investigation to understand its influence separate from bandwidth (Hawks & Miller, 1995; House 1960). Given these open questions, the current study aims to analyse amplitude and bandwidth in conjunction. Since abnormal reductions in bandwidth impact naturalness despite minimal perceptual shifts (Remez et al., 1981; Stevens et al., 1969), assessing amplitude's interaction can uncover if similar patterns emerge. By taking a broader spectral focus encompassing multiple parameters, a clearer model of vowel production mechanics may develop. The goal is not claiming any one measure as primary, but rather elucidating their relative contributions and interdependencies in signalling contrasts. Findings can guide future studies exploring if certain cue weightings hold true cross-linguistically or prove more language-specific. Ultimately, this research intends to add another dimension to the complex puzzle of vowel perception.

Taking this into account, the rationale for selecting measures is threefold. Firstly, both spectral moments and measures have received limited attention in speaker comparison research, as previously noted. Secondly, no existing work has evaluated the combined potential of these two feature types for speaker comparison. Lastly, since these features MAY individually exhibit speaker-specific traits, it is hypothesised that assessing them alongside centre formant frequencies could enhance speaker classification system performance.

This study will systematically investigate these reasons, examining the features individually, jointly and in diverse contexts to ascertain their efficacy as speaker discriminants.

2.4 Research Questions and Motivation Behind the Study

As mentioned in Chapter 1, the study raises an important question regarding the focus on labour-intensive and time-consuming aspects of speaker comparison rather than utilising advanced ASR systems. The first rationale provided for this methodological choice was that there is limited existing research on specific features, such as within-formant features, which could carry speaker-specific characteristics. By examining them in detail, the study aims to enhance system efficiency, along with centre formant frequencies. Secondly, the research concentrated on interpretable spectral features to improve transparency and reliability in contrast to the opaque decision-making processes of ASR systems. Finally, the study acknowledges the complex admissibility of ASR-derived evidence in various legal contexts, emphasising the need for a human-assisted interpretable system to address legal challenges and varying acceptance standards worldwide. Overall, acoustic-based techniques allow for

comprehensive analyses of fine-grained speaker traits while providing transparency for legal applications.

As the study aims to investigate all the reasons mentioned above by analysing within-formant features in forensic casework, the following research questions were formulated for each article:

2.4.1 Article 1

1. Can a spectral moment analysis (SMA) of the four moments of vowel formants F1–F4 help distinguish between individual speakers?
2. Are there factors that either impede or facilitate the discriminant values of spectral moments (SMs)? If so, what are these factors?

In that regard, we may ask,

- 2.1 Which SMs and combinations of SMs are most effective?
- 2.2 Which vowels or subsets of vowels show the best discriminant value?
- 2.3 Which elicitation techniques and the associated speaking styles provide the best data for SMA?
- 2.4 Which varieties of Marwari does SMA work best on?

2.4.2 Article 2

1. Are spectral measure values impacted by variety, vowel and mode of data elicitation?
2. Can including spectral measures with formant centre frequencies help distinguish between individual speakers in an acoustic analysis?

If yes,

- 2.1 Which spectral measures and combinations of spectral measures are most effective?

3. Are there any factors that impede or facilitate spectral measures' discriminant values?

If so, we may ask,

3.1 Which vowels or subsets of vowels yield the highest classification rate (CR) results when spectral measures are applied?

3.2 Which speech styles provide the highest CRs when spectral measures are applied to them?

3.3 Which varieties do spectral measure analysis best work on?

2.4.3 Article 3

1. Which feature clusters from the F1–F4 centre formant frequency range have the highest speaker discriminatory power: bandwidth, within-formant skew, within-formant kurtosis of energy, formant amplitude, relative amplitude centre of gravity, standard deviation, or spectral peak?
2. Does combining within-formant spectral moments (centre of gravity, standard deviation, skewness, and kurtosis) and spectral measures (formant amplitude, formant bandwidths, and spectral peaks) improve the accuracy of speaker classification?

   If so:

   2.1 Which spectral feature combination has the greatest speaker discriminatory value?

   2.2 Which vowels and vowel subsets have a greater discriminatory value for spectral feature analysis?

   2.3 Do spectral features or feature combinations perform better for some modes of data elicitation than others as speaker discriminatory features?

   2.4 Does the speaker discriminatory power of spectral feature analysis improve for some varieties more than others?

The subsequent chapter will introduce the Marwari language.

3 Marwari Language

This chapter provides a general overview of the Marwari Language and its speakers. The chapter begins with a brief geographical and historical overview of the language. This introduction is followed by an outline of the Marwari's political and social landscape. This section concludes with a brief introduction of the Marwari language's power dynamics in comparison to the other dominant language in the area, i.e., Hindi.

The second section of this chapter presents a brief linguistic outline of the Marwari language. This section is broken into three sections; The first exhibits the broad linguistic aspects of the language. The second section concentrates on the phonology of the Marwari language. The final component discusses the varieties employed in the current investigation. This section discusses the genesis, hierarchical significance, and fundamental distinctions between these kinds.

3.1 Geographical Distribution of the Language

Marwari is an Indo-Aryan language, spoken primarily by the members of the Marwari community (also called "Marvari," "Marvadi" and "Marwadi"). People from the Marwari community have been residing in the north-western areas of Rajasthan (a north-western state of India), notably Jodhpur, Bikaner, Barmer, Nagaur, Pali, and other neighbouring districts. Rajasthan is located in the northwest of India and has an area of 342,239 square kilometres (Office of the Registrar General & Census Commissioner, 2011c). This language is also spoken in the neighbouring state of Gujarat and in some regions of Pakistan. In addition to different languages/varieties that fall under the Rajasthani umbrella, Marwari is also surrounded by Hindi and its various dialects (Mukherjee, 2011). She also mentions that Marwari gradually assimilates with "Gujrati" through the "Bhili/Bhiodi" along the east-to-west territory and meets with Sindhi, Lahnda and Punjabi in the northern zone.

The next section will offer a brief outline of the Marwari speaking population.

3.2 Population of Marwari Speakers

According to the 2011 census report, the population of Rajasthan is around 68,548,437 people, including 7,831,749 Marwari speakers, which is more than double the figure from the 1991 census report. (Samuvel et al., 2012). The Marwari language is sometimes grouped with other

languages spoken in the state of Rajasthan, such as “Haroti” and “Mewadi.” According to Language (2011), 25,806,344 people reported that their mother tongue is Rajasthani, which, together with those identifying Hindi as their mother tongue, suggests that the real number of Marwari speakers is more than what was recorded in the census. Shougrakpam (2022) puts this number between 45 to 50 million. This misrepresentation is often associated with the prestige and geo-political association of the language. The Marwari language is regarded as prestigious among the various languages spoken in Rajasthan, because it is frequently used in media, literature and education. However, a sociolinguistic survey of selected Rajasthani speech varieties revealed that the speakers of many Rajasthani dialects, including Marwari, report Hindi or Rajasthani as their mother tongue (SIL Electronic Survey Reports, 2012). This might be due to prestige, confusion over what is “Rajasthani” against “Marwari” and/or speakers’ lack of knowledge about what distinguishes their variety as a “dialect” from a “language” (Shougrakpam, 2022). The important point to note here is that there is no language called “Rajasthani,” and the term “Marwari,” which is usually linked with the language spoken in Rajasthan’s Marwar region, may have more speakers than reported by any official government or academic source claim (Samuvel et al., 2012; Shougrakpam, 2022).

As mentioned in the previous paragraph, the distinction between language and dialect is not always clear for Marwari language speakers when reporting to the census or any survey. It is critical that this distinction should be addressed before continuing with the current investigation. For sociolinguists, the distinction between language and dialect remains a point of contention. The same challenge persists in the classification of Marwari. There has been an ongoing struggle among Marwari speakers to establish it as a “language” rather than a “dialect” or a “variety” of Hindi. As stated in Lewis and Summer Institute of Linguistics (2009):

> “Every language is characterised by variation within the speech community that uses it. Those varieties, in turn, are more or less divergent from one another. These divergent varieties are often referred to as dialects. They may be distinct enough to be considered separate languages or sufficiently similar to be considered merely characteristic of a particular geographic region or social grouping within the speech community. Often speakers may be very aware of dialect variation and label a specific dialect with a name. In other cases, the variation may be largely unnoticed or overlooked (p.05).”

To avoid additional misunderstanding, Marwari will be treated as a distinct "language" in the present study, and a more neutral word "variety" will be used for different caste-based analyses.

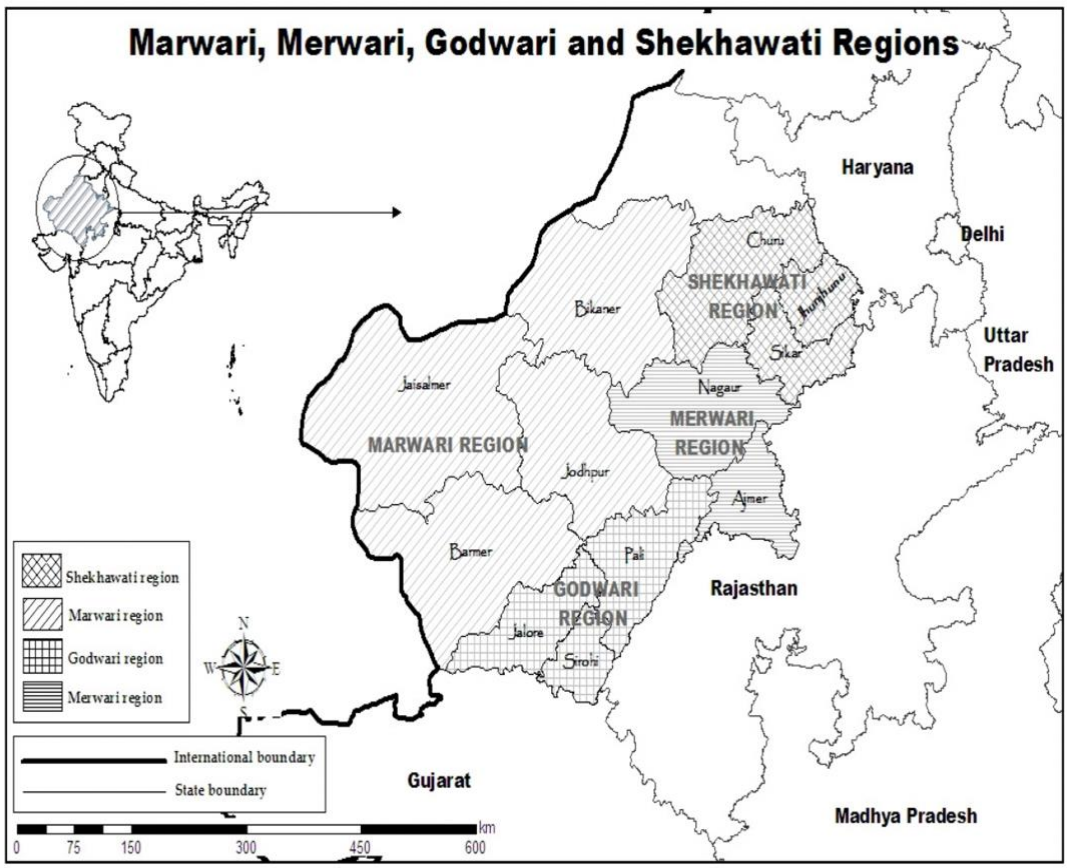


Figure 3.1 Map of Rajasthan depicting the geographical boundaries of the Marwari-speaking area (Chacko & Ngwazah, 2012).

The next section will provide a brief overview of the current political and social standing of the Marwari language.

3.3 Current Political and Social Standing of Marwari

Beshears (2017) quotes a saying in the Marwari language that goes, "If you go three kilometres, the taste of water changes. Go six kilometres, and the language changes" (p. 04). This description of the language is applicable to the entire Indian subcontinent. In a country whose state boundaries are established based on linguistic disparities, there are numerous unrecorded linguistic variations within each area. It is challenging to define a definite boundary of any given language in the country, considering it changes according to caste, clan and region.

It has been a huge challenge for the multilingual government to find a definitive answer to these differences (Liu, 2023). The implementation of the three-language policy in 1968 as an educational reform is one approach the administration chose to overcome this barrier (Meganathan, 2011). This three-language policy mandates that non-Hindi speaking states teach Hindi, English and a regional language (Bhattacharya, 2017). In Hindi-speaking states, students would be taught Hindi, English and another regional language (preferably from south Indian states). The instruction medium could be any one of these languages (depending on the school's system). As a result of this change, most Hindi-speaking states now use either Hindi or English as the medium of instruction. In a non-Hindi state, either the regional language or English are the preferred medium of instruction. The policy also proposed that the primary education mode can be the student's mother tongue, with Hindi and English serving as second languages. The main issue with this approach is that children may only choose from a list of scheduled languages as their mother tongue, and many "languages," including Marwari, are not on this list. The use of any other mother tongue is often disapproved of or completely banned (Alekseevna & Sergeevna, 2021).

The original three-language policy went through many changes over time (Batra, 2020; Bhattacharya, 2017). The most recent change was the education reform of 2020, which removed the mandatory Hindi and English implementation from the law (Aithal & Aithal, 2020). Now, the students can select any language (regional or foreign) they want to study from the government's scheduled language list. As mentioned earlier, the list only included 22 out of 19569 mother tongue languages spoken in the country. This data was based on languages that had more than 10,000 speakers (Abbi, 2010). According to Chandramouli (2011) this socially accepted language, with officially having 78,31,749 speakers, is still struggling to be recognised as an independent language and is still considered a dialect or a variety of either Hindi or Rajasthani languages. Hindi as a dominant language continuously creates a hierarchical image among the Marwari speakers that results in the speakers identifying themselves as Hindi speaker rather than Marwari speakers (Mukherjee, 2011).

Religious variations have a significant effect on the diversity of most north-Indian languages. However, for the Marwari language, this process becomes much more problematic because in addition to religious distinctions, caste and area also play an important part in the society (Office of the Registrar General & Census Commissioner, 2001). The geographical differences in the Marwari language are particularly noticeable. Because in order to the severe travelling

circumstances caused by the Thar desert, most regional differences survived until the eighteenth century (Kothiyal, 2021). People did not migrate as often in these areas as they did in the rest of the country. Finding literature related to migration to or from this area before the partition of India is very difficult (Kothiyal, 2016). The Census of India (1911) talks about migration a little with its primary focus being the migration within the Thar area or neighbouring states (Bikaner was a princely state of Rajputana). This kind of migration was only temporary and often corresponded with seasonal work (Khera, 2005). Nakatani (2017) mentions that in the nineteenth century, there was a small migration of traders and bankers of the Marwari community within the country. The study mentions that it was not until the twentieth century that people started migrating to the port cities of this community. Though this migration was a significant change for the community, people who migrated belonged to a tiny group and stayed very close. The term "Marwari" became a synonym for people from the business class outside Rajasthan (Roy, 2015). This kind of focused migration kept the language of the Marwaris intact. Similarly, the caste system's strict code that stopped people from marrying, interacting, or mixing among the castes ensured the sustainability of various caste dialects within the same region.

The Marwari-speaking community is primarily a business community. The post-independent social laws and the lack of proper resources limited their access to travel (Magier, 1983; Nakatani, 2017). The new regulations and the transfer of European industrial firms made the previous trader industrialist (Roy, 2015). As a result, migration between Marwari communities rose significantly after independence, and most Marwari speakers became bilingual. Finding monolinguals among the younger and more urban Marwari speakers is quite difficult. Watson (2017) explored the evolution of Marwari language attitudes across generations. Her research provides a detailed summary of how the previously monolingual Marwari speech community became bilingual for various socioeconomic reasons. People in their 60s or older were the only remaining monolinguals in the neighbourhood. Because of a lack of a robust education system and traditional social beliefs, the majority of the rural women population in the "Marwar" region also stayed monolingual (Watson, 2017).

Hindi is a compulsory language in most marketplaces, educational institutions, and government offices. Because of their limited connections with the outside world, those who reside in rural areas might evade this forced Hindi impact. To eliminate any Hindi impact on the participants, women from rural and semi-urban communities from three distinct castes were chosen for the

current experiment. These caste varieties were picked at random. Further detail on the selection process will be provided in Chapter 4.

The following section tries to describe the linguistic background of the Marwari language spoken in Bikaner. The section will also provide a general overview of the varieties selected for the current project.

3.4 Linguistic Overview of Marwari

Linguistically, Marwari belongs to the Central group of the Inner Indo-Aryan Language family. As mentioned earlier, Marwari has been grouped under Rajasthani as a mother tongue in Indian Census Reports till 1961, and since 1971 it has been considered one of the dialects or varieties of the Hindi (Mukherjee, 2011; Office of the Registrar General & Census Commissioner, 2001, 2011a, 2011c). Nigam (1972) has defined Marwari as "the principal dialect of Western Rajasthan spread along a wide area and written as a common form of speech by the native speakers who are found spread all over the country" (pp. 162-163). Marwari is also known to have several traits showing affinity with Sindhi (Mukherjee, 2011). The mutual intelligibility of most neighbouring dialects of Marwari also makes it confusing to create a clear boundary between these languages (Shougrakpam, 2022). To understand this in further detail, the following figure tries to provide a derivational status of Indo-Iranian linguistics. The word Marwari appears here as a part of the Central Indo-Aryan Languages.

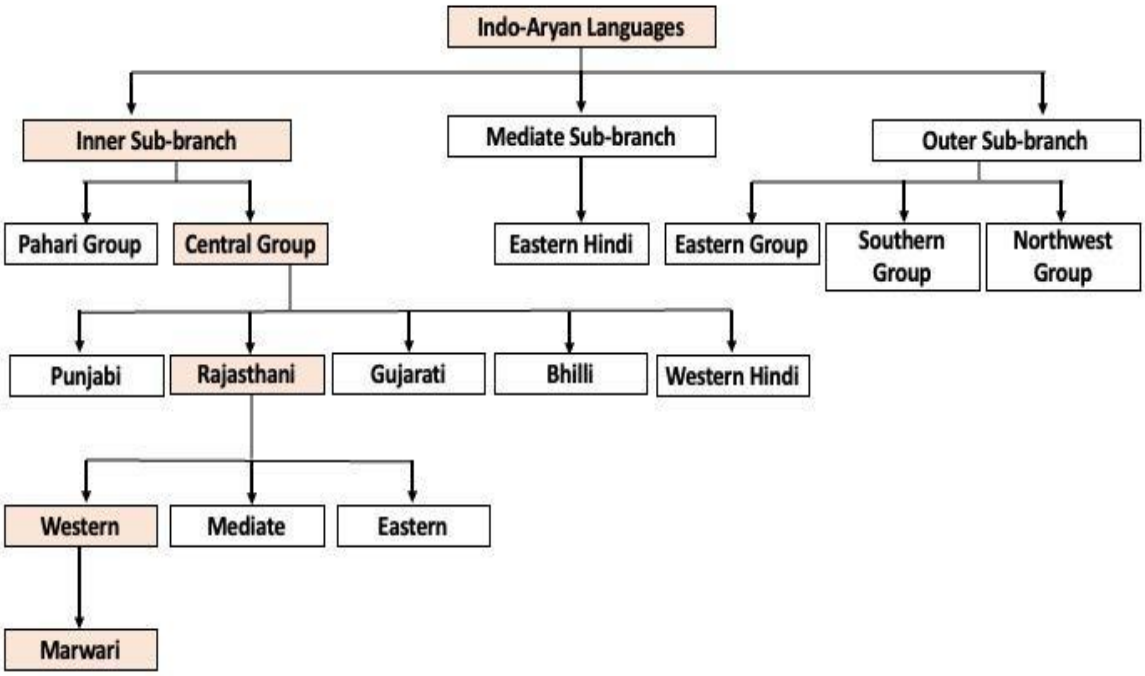


Figure 3.2 Classification of Indo-European languages (Adapted from Masica, 1991, p. 449).

3.4.1 Grammar of Marwari

Until recently, Marwari was mostly a spoken language; most ancient knowledge of the language still exists in oral forms (Kothiyal, 2016). These oral traditions were preserved through folk songs and storytelling. There is no official written script for Marwari, however, the "Devanagari" script has been approved with minor adjustments in India, and the "Perso-Arabic" script has been accepted in Pakistan (Shougrakpam, 2022). The lack of a written form of the language also indicates that until recently, there was no formal or prescriptive grammar. Marwari norms and traditions were passed down orally for a very long period. This oral literature began to appear in written form in the seventeenth century. (Ziegler, 1976). There is still a lot of different and misleading literature available on grammar. This could be associated with the fact that the language also differs from region to region (Phillips, 2012). An extensive study is required to assess the dialect continuum of the language.

Marwari language, as most of its sister languages, follows Subject-Object-Verb (SOV) word order. The overtly marked case makes the sentence structure very versatile, i.e., the order of sentences can be changed depending on the context or need. Marwari has grammatical gender

and number, i.e., both features are marked overtly on the verb. There are two grammatical genders in Marwari -- masculine and feminine. Marwari follows a very strict gender, number, and case agreement on the verb. There are 11 cases (Nominative, Agentive, Benefactive, Accusative, Instrumental, Dative, Ablative, Genitive, Locative, Sociative and Vocative) (Mukherjee, 2011). Marwari is a highly inflectional language where all three of these features could be marked on a single word (sometimes as a single stem). The verb can agree with both subjects and objects (Subject-verb agreement in most cases, and Object-Verb agreement in Ergative case). The following section provides a brief overview of some other grammatical features. The data presented here is the Researcher's native intuition of the language from the Bikaner District. Some abbreviations used in the following sections are:

1st Person: - 1	Auxiliary: - Aux	Preposition: - Prep
2nd Person: - 2	Verb: - V	Present Tense: -Pres
3rd Person: - 3	Negative: - Neg	Past Tense: - Past
Singular: - S	Masculine: - M	Future Tense: -Fut
Plural: - P	Feminine: - F	Honorific Marker: - Hon

Word order: SOV (Subject- Object- Verb) Being a highly inflectional language makes it possible for the speakers to alter the word orders if needed.

sɪtɑ	**ɪskʊl** [4]	**dʒɑv-ɛ**
Sita 3S.F	school	go. V. Pres. 3
(Subject)	(Object)	
sɪtɑ ɪskʊl jɑvɛ		
Sita goes to school.		

[4] School is one of the many English loan words that have become part of Marwari's core vocabulary overtime, however, an additional vowel is inserted either before or between the consonant cluster 'sch' in school.

The preposition follows the noun. The preposition /pər/ is following the noun it is representing.

peɳ	**mɛdʒ**	**pər**	**ɖʰər-o-ɽo**	**hɛ**
pen[5] M	table F	On Prep.	Keep.V.Pres.3SM	be. Aux. Pres
(Subject)	(Object)			
peɳ mɛʒ pər ɖʰəroɽo hɛ				
the pen is kept on the table				

An adjective precedes the noun as a head. The below example shows that the adjective good is preceding the noun head pen.

tʃʰokʰ-o	**peɳ**	**mɛdʒ**	**ɖʰəroɽo**	**hɛ**
good. S. M	pen M	table F	keep.V.Pres.3SM	be. Aux.Pres
	(Subject)	(Object)		
çʰokʰo peɳ mɛʒ pər ɖʰəroɽo hɛ				
A good pen is kept on the table				

The adverb always precedes the verb.

kʰɑt-o	**tʃɑl**
fast. S. M	walk.inf
kʰɑto çɑl	
Walk fast	

[5] Pen is one of the many English loan words that have become part of Marwari's core vocabulary overtime.

Verb is followed by auxiliary.

peɳ	**mɛdʒ**	**pər**	**ɖʰər-o-ɽo**	**hɛ**
pen M	table F	On Prep.	Keep.V.Pres.3SM	be. Aux. Pres
(Subject)	(Object)			

peɳ mɛʒ pər ɖʰəroɽo hɛ

the pen is kept on the table

The negative proceeds the main verb

peɳ	**mɛdʒ**	**pər**	**ɖʰər-o-ɽo**	**koni**
pen M	table F	On Prep.	Keep.V.Pres.3SM	No. Neg
(Subject)	(Object)			

peɳ mɛʒ pər ɖʰəroɽo koni

the pen is not kept on the table

The indirect object precedes the direct object.

mʰʊ	**peɳ**	**mɛdʒ**	**pər**	**ɖʰər-o-ɽjo**	**koni**
I.1S.M	pen M	table F	On Prep.	keep.V. Past.3SM	No. Neg
(Subject)	(Indirect Object)	(Direct Object)			

mʰʊ peɳ mɛʒ pər ɖʰəroɽjo koni

I did not keep the pen on the table

It is a ‘pro-drop language. i.e., in a compound sentence the subject can be omitted in the second phrase.

miːrɑ	**ɑɪ**	**r**	**ɑpɑ**	**sɑgɛ**	**dʒiːmsɪ**
Mira	come.3SF. Pst	and	we	together	eat.V.3S

miːrɑ ɑɪ r ɑpɑ sɑgɛ ɹiːmsɪ

Meera will come and eat with us.

Verb shows the tense, aspect, and mood markings.

tʃʰoro	**rəm**	**lɪo**
boy.3.S.M	play. V	to take.V.3SM

çʰoro rəm lɪo

The boy has played

Marwari has grammatical numbers and gender.

peɳ	**mɛdʒ**	**pər**	**ɖʰər-o-ɽo**	**hɛ**
pen M	table F	On Prep.	keep.V.Pres.3SM	be. Aux.Pres
(Subject)	(Object)			

peɳ mɛʒ pər ɖʰəroɽo hɛ

the pen is kept on the table

The language has a complex honorific system.

ʈʰɑr-ɑ	**dɑdʒɪ**	**kɑl**	**ɑsɪ**
You.3. P.M. Hon. Pos	grandfather	tomorrow	come. Fut.Hon

ʈʰɑrɑ dɑzɪ kɑl ɑsɪ

Your grandfather will come tomorrow

3.4.2 Phonology of Marwari

3.4.2.1 Marwari Vowels

Magier (1983) presented an overview of Marwari phonology, noting similarities to Hindi vowels but with one key difference: the mid vowels are less clearly distinguishable. His study

thoroughly catalogued the vowel inventory. According to him, Marwari has ten vowels with no contrastive lip rounding - /i/ high front, /ɪ/ lower high front, /e/ mid front, /ɛ/ low front, /ə/ mid central, /ə/ low central, /u/ high back, /ʊ/ lower high back, /o/ mid back, and /a/ low back. While back vowels are rounded, rounding is not phonemic in Marwari. Out of these vowels, /ʊ/, /ə/, and /ɪ/ are short while the rest are long. Magier also observed that dialectal differences result in vowel mergers, especially involving the mid vowels, where the higher-mid and lower-mid vowels often seem to collapse in the spoken form for many speakers.

Gusain (2004) confirmed previous findings, citing the same vowel inventory of four front vowels, two centre vowels, and four rear vowels. The most recent examination by Mukherjee (2013) shows the same vowel inventory, indicating that there has been no significant vowel change since Magier (1983). Table 3.1 summarises the vowels recorded in all three investigations.

Table 3.1 Vowel inventory of Marwari language (first published in Magier, 1983 and later adapted in Gusain, 2004)

	Front	Near-front	Central	Near-back	Back
Close	iː				uː
Near-close		ɪ		ʊ	
Close-mid	e				
Mid			ə		o
Open-mid	ɛ				ɔ
Near-open					
Open			ɑː		

3.4.2.1.1 Vowel Nasalization and Length in Marwari

All three referenced investigations (Magier 1983; Gusain 2004; Mukherjee 2013) determined that vowels in Marwari can be nasalized phonetically but only word or syllable-final nasalization before a pause is phonemic, as in examples like /ã ĩ ẽ õ ũ/. Additionally, non-final oral vowels become nasalized before nasal consonants, evidenced in forms such as /kãʈ/ ‘thorns’ and /sĩːg/ ‘horn.’

The studies also concur that vowel length is contrastive in Marwari, with short and long counterparts distinguishing meanings as validated through minimal pairs across word positions. Although Magier (1983) posits it as a feature differentiating Marwari vowels, he does not proide supporting examples. Gusain (2004) however substantiates this analysis by citing pairs such as:

/sɪl/ ‘stone slab’ /siːl/ ‘damp’

/ʊmra/ ‘king’ /uːmra/ ‘raw crop’

Additional instances are provided in Mukherjee (2013):

/ɖɪn/ ‘day’ /ɖiːn/ ‘poor’

/dhʊn/ ‘tune’ /dhuːn/ ‘concentration’

3.4.2.1.2 Vowel Gliding and Harmony Patterns

Gliding of vowels occurs before palatal and labial glides resulting in diphthongs. For example, /miɛl/ ‘dirt’ and /dɦau̯l/ ‘running’ (Magier 1983).

Backness harmony is also observed where suffixes take on the backness specification of stem vowels, as with /ɖhũːk-ko/ ‘hill’ and /bĩʈ-ke/ ‘wall’ showing /u/ and /i/ conditioning respectively (Magier 1983).

3.4.2.1.3 Tone and Vowel Length

Both Magier (1983) and Gussain have noted the existence of lexical tone contrasts in Marwari analogous to Punjabi, with three identified tones - high, mid and low.

According to Magier (1983) a tonal contrast in some forms like /kər/ ‘do’ versus /kər-iː/ ‘will do’, marking it as a high-falling tone that can occur with any vowel. Short vowels exhibit lengthening in such future tense forms in addition to the tone.

The high tone (◌́) manifests as a rising tone. The low tone (◌̀) is characterized by a falling contour. The mid tone (◌̄) is predictable by redundancy rules and is not overtly marked, as vowels without a specified tone carry this by default. Unlike tonal languages such as Chinese having contour tones, Marwari only exhibits level tones on syllables.

Some examples reflecting the phonetic nature of the level tones provided in Gusain (2004) are:

/pèr/ ‘duration’ /pér/ ‘leg’

/chə̀r/ ‘put on’ /char/ ‘wave’

/lèr/ ‘behind’ /lìr/ ‘taken’ (past participle)

/kèr/ ‘calamity’ /kér/ ‘caparis’ /kə́r/ ‘said’ (conjunctive participle)

/nàr/ “having bathed” /nár/ “women” /nâr/ ‘tiger’ (p.16)

In this representation, tone markings are placed over the syllable carrying the tone. The consonant onsets do not carry the tone. Mukherjee did not mention any tonal contrast in the language. It should be noted that the evidence presented here needs to be further examined for confirmation.

While Gusain and Magier highlight its presence across the wider Marwar region, Mukherjee's Bikaner-focused study doesn't mention it. This disparity suggests two possibilities: first, regional variations, similar to Punjabi, Marwari's tone might be fading, particularly in Bikaner, leading to its absence in Mukherjee's data. And/or Gusain and Magier might have prioritized aspects more broadly present in Marwar, including tone ,while Mukherjee concentrated on Bikaner-specific features, potentially overlooking tone.

3.4.2.2 Marwari Consonant System

The Marwari stop system shows a 4-way contrast between voiceless unaspirated /p t̪ ʈ tʃ k/, voiceless aspirated /ph t̪h ʈh tʃh kh/, voiced unaspirated /b d̪ ɖ dʒ g/, voiced aspirated /bɦ d̪ɦ ɖɦ dʒɦ gɦ/. Examples: /phoʈ/ 'explosion', /poʈ/ 'bandage', /boʈ/ 'descendant', /bɦoʈ/ 'bowl' (Magier, 1983).

Voiced stops are phonetically realized as implosives [ɓ ɗ̠ ɗ̠ɦ]. Voiceless unaspirated stops become voiced between vowels: /pak-na/ → [pəgnə] 'to ripen' (Magier, 1983).

Affricates /ʧh/, /dʒ/, /dʒɦ/ also occur. Retroflex /ʈ ɖ ɽ/ contrast with dentals /t d r/. All stops/affricates distinguished for nasals too (Magier, 1983).

Approximants /j/ and central /ɥ/ occur. No phonemic fricative contrasts, only allophonic (Magier, 1983).

In Marwari two aspirated cannot occur together in a word. All voiced stops are generally implosives. In Marwari, nasalisation patterns lead to changes in vowel quality of final vowels.

Table 3.2 Phonemic inventory of the Marwari consonants (Mukherjee, 2011)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Glottal
Plosive	p b		t̪ d̪			ʈ ɖ		k g	
Plosive (Aspirated)	pʰ bʰ		t̪ʰ d̪ʰ			ʈʰ ɖʰ		kʰ gʰ	
Nasal	m			n			ɲ	ŋ	
Trill				r					
Tap or Flap						ɽ			
Fricative		f v		s	ʃ				h
Lateral fricative									
Approximant							j		
Lateral approximant				l		ɭ			
Affricate					tʃ dʒ				
Affricate (Aspirated)					tʃʰ dʒʰ				

Note: Unlike English, all the stops and affricates are in contrastive distribution with their respective aspirated version, and therefore they are separate phonemes.

Syllable structure in Marwari is predominantly CVC, with a maximum of four syllables per word (Mukherjee, 2011). However, Bikaneri Marwari exhibits a distinctive aversion to consonant clusters within syllables. To avoid clusters, speakers often insert vowels, mainly /ɪ/ or /ə/, between consonants. This insertion appears largely arbitrary, as evidenced by borrowed Hindi words: “stʊ-tɪ” (prayer) might become “sə-tʊ-tɪ” or “ɪs-tʊ-tɪ” in Bikaneri speech.

The stress in Marwari is both contrastive and emphasis based. The stress falls on the first syllable for nouns and the second syllable for verbs in two-syllable words. The following figure shows the contrastive difference observed in three different nouns and verbs of the Marwari language.

Table 3.3 The noun-verb contrastive stress pattern in Marwari language (Gusain, 2004).

Nouns		**Verbs**	
ˈ**bəlla**	Evil spirit	bə ˈlla	Call
ˈ**hilla**	Job	hi ˈlla	Cause to move
ˈ**silla**	Rumour	si ˈlla	Cause to wet

This can also be affected by the length of the vowel, or the emphasis required on the word. The present study only looked at the stressed vowels from conversation and story data. See chapter 5 (section 5.1) for further details.

3.5 Within-Language Variability

The caste system in India is an ancient system of differentiating social class according to individuals' work. As mentioned earlier in section 3.1, Marwari, like most of the Indo-Aryan languages spoken in Northern India, has different varieties depending on class, region, caste and religion. The current study examines caste variations in Marwari by accessing informants across different strata. Caste selection was governed by participant availability and researcher contacts rather than comprehensive enumeration. The three caste groups represented were Brahmin, Bishnoi and Jaat. While inclusivity across caste hierarchy was intended, the caste list is not exhaustive for the Marwari speech community. Participant recruitment relied on researcher accessibility within personal acquaintanceship circles spanning the highlighted castes. To account for geographical variance, the study is limited to caste members residing in Bikaner. This section will offer a general overview of the selected castes as well as a quick description of the caste dynamics in Bikaner. The following section will provide a brief overview on the geographical background of the region.

3.5.1 Geographical variation

Rao Bikaji established the city of Bikaner in 1488 CE, transforming the Thar Desert territory formerly known as Jungladesh into a metropolis (Census of India, 1911). The city has a wall

around the old town with gates that used to close in the evenings. Historically, sections were demarcated by caste, as colony names reflect today despite recent shifts. This segregation enabled preservation of distinct dialectical characteristics tied to caste identity. Another reason for caste-based dialect preservation could be that, for generations, Marwari was the primary desert language. Limited mobility and caste endogamy-maintained isolation of linguistic features. However, post-1947 substantial migration increased inter-caste contact, dramatically shifting boundaries as Marwari varieties mixed (Nakatani, 2017). This aligns with Wolfram's (1997) idea of a post-insular language situation, characterised by historically isolated language varieties transitioning out of seclusion as a result of wider interaction with speakers from other groups or different demographic changes such as population movements (p.3).

Figure 3.3 Position of Bikaner district in Rajasthan (Google Maps, 2023).

Yet due to the strong identification of its populations with their dialects, Marwari, spoken by approximately 22 million people in north-western India, retained many ancient dialectal divides. These distinctive characteristics linked to dialects are more prominent among women in these areas. One reason could be that due to the presence of patriarchal and caste-specific practices, migration was largely undertaken by males. Given this cultural context, female

subjects from the Bikaner region were chosen for the study in order to minimise the influence of other languages (Nakatani, 2017).

These females were selected from three different castes. The intricacy of Hindu social hierarchy makes elucidating caste complex. Still, examining minute linguistic distinctions across caste varieties can further knowledge without aiming to spark controversy over varna rankings. The goal is to determine whether these minute linguistic distinctions offered by each caste-based variety may aid in the current study.

The following sections will provide a brief background on the caste system with a summary of the castes selected for the study.

3.5.2 Caste

Before examining the variety specific-variations, it is helpful to differentiate the social dimensions upon which these varieties are founded. Numerous research studies have been conducted to investigate dialectal and social distinctions within language (e.g. Barber et al., 2012; Kerswill, 2003). The majority of these distinctions are based on social class. In sociolinguistics, caste has been shown to correlate with linguistic variation similarly to social class across Indian languages. Though geographic, gender, and individual variation can also play a role, historical data suggests caste remains an influential factor (Mukherjee, 2011). The study quantitatively investigates phonetic distinctions between the three caste varieties, which have not previously been studied acoustically. This establishes a framework for applying insights to speaker identification and forensic purposes in future research. As background, this section will provide a brief overview of what caste is while also summarise the historical context of the select castes.

The origin of the word ‘Caste’ comes from the Portuguese word ‘Casta’, meaning something that ‘cannot be mixed’, or ‘pure’ (Saha, 1993). The caste system in the Hindu religion is an occupation-based social division of people into various groups. People who share the same caste share a common ancestry and, in most cases, the same last names (Béteille, 1967).

Before digging into the notion of caste, it is necessary to comprehend the concept of “Varna.” The Hindu social system is divided into four varnas (Burghart, 1978). These varna represent the body of ‘Brahma’ (creator of the universe). They define varna as a term that represents the colour and the derivation rank (Burghart, 1978, p.521), or the four parts of Brahma’s body that

work together to keep the universe running. Each varna is based on their ability to execute a duty in society, and there is no hierarchy since competence is measured not by the nature of the job, but by the virtue of completing the tasks of their own varna. However, traditionally, each of these varna was connected with a social order based on the purity of their work. These four parts are (Burghart, 1978):

Brahmins: The highest varna are "Brahmins." Brahmins are the brain of "Brahma." Their primary role in this system involves education, priesthood, and knowledge.

Kshatriyas: "The Warrior Varna" or "Kshatriyas" dedicate their lives to rule and perform their caste duties as warriors. They are the arms of "Brahma."

Vaisya: Their duty is to herd cattle, farm and provide food to everyone. There are the thighs of "Brahma." Most Vaisya's are involved in commercial activities of the society, for example, business and construction.

Sudra: This varna came from the feet of "Brahma." Their archaic function was to serve the rest of varnas.

As the varna reflects the social hierarchy, there may be multiple castes within each varna, and within that hierarchy. For example, the castes "Suthar" (carpenters), "Luhar" (blacksmiths), and "Sunar" (goldsmiths) all belong to the same varna because they do a variety of intricate activities that serve society. It should be mentioned that the Indian constitution eliminated caste hierarchies, which should not be practised in principle. But, even today, each caste has its own social position, linguistic variety, and set of regulations, which means that each caste has its own variation of the Marwari language. This variation is often based on phonemic differences (including different rules), different choices of vocabulary. However, the syntactic structure and grammar of these varieties remain the same. The areal differentiation typically assisted them in maintaining these distinctions. Although these biases have been formally erased in modern-day Bikaner, individuals continue to strive to retain their identities through their distinct linguistic dialects. Saha (1993) examined the significance of caste and its associations with a specific variety of a language in relation to the administration of occupations in society. Every dialect is mutually intelligible and often very similar to the other varieties. The current study will focus on three significant castes in Bikaner based on the availability and access to the participants: Jaat, Bishnoi and Brahmin. Brahmins belong to the highest level of the Hindu caste system. Jaat belonged to the Vaisya varna, the third level of the caste system. The Bishnoi

caste is descended from the Bishnoi community, which was formed to eliminate castes but eventually evolved into a caste itself (see section 3.5.3 for further details). However, their place in the caste structure varies in each state, for a variety of geopolitical reasons. The following section provides a summary of these castes.

3.5.2.1 Brahmin caste

Brahmin, as mentioned earlier, is the highest varna of the Hindu caste system. Their superiority derives from the power that their occupation holds as a link between Gods and humans (Saha, 1993). The term Brahmin, unlike any other varna, is also used to mark the caste of the people from this varna. Brahmin caste is subdivided into various categories. For the current study, the Marwari spoken by the "Pushkarna" Brahmins was selected. The naming within the Brahmin caste categories is mainly based on the origin of their respective ancestors. Pushkarna Brahmins originated from the small town of Pushkar. The earliest mention of Brahmin caste is found in 925AD in an inscription (Jain, 1979). Pushkarna Brahmins chosen for this study are all the descendants of Brahmins who moved to the city at least three generations ago.

3.5.2.2 Jaat caste

The Jaat caste has its historical roots in Northern India around the Indus Valley. Originally pastoralists, Jaat (or Jat) migrated across the northern Indian plains for centuries as early as the one BC (Nijjar, 2008, p. 44). They have been a very prominent community politically and socially in Rajasthan since the 17th century. The community consists of 9.2 per cent of the population of the state of Rajasthan (Rathore & Saxena, 1987). Despite being one of the most prominent communities of the state, the people primarily reside in the rural or suburban areas of Rajasthan. Though the Jaat community in other states falls under the Kshatriya varna (Punjab or Haryana) in Rajasthan, Jaat belongs to the Vaisya or Shudra varna of the Hindu caste hierarchy based on their occupation as business, farmers and cattle herders.

3.5.2.3 Bishnoi caste

The Bishnoi caste, or most appropriately community (the term community is more appropriate as this community was formed to abolish caste hierarchy), is the newest caste among the three castes selected for this study. People started following Jambheshwar Ji in 1485, with their identity primarily defined by their shared faith and adherence to the 29 principles (Jain,

2010). This emphasis on religion transcended traditional caste divisions and created a sense of unity among Bishnois (Sinha & Singh, 2020).

Originally, within the Bishnoi community, social hierarchy was less rigid compared to the caste system. While there might have been some differences in occupation and landownership, these were not determined by birth but by individual circumstances and merit (Jain, 2016). However, it's important to note that the relationship between caste and community in India is complex and nuanced. While the Bishnoi may not strictly adhere to the caste system, they are still embedded within the broader social context of India, where caste continues to play a significant role. Some scholars argue that the Bishnoi community can be seen as a "neo-caste" or a "sect-caste," as it exhibits some characteristics of both castes and communities (Kavoori, 2002).

A lot of the present Bishnoi speakers have their roots in the Jaat community before the 15th century. Their respective cultural and social values are very similar to each other. Bishnoi speakers' distribution is mostly around India's north-western regions, with a significant concentration in Haryana and Rajasthan (Jain, 2010). Phonetic differences and similarities among the selected castes:

Phonetically, caste-based varieties of Marwari show some linguistic differences, as exemplified by variations in pronunciation of certain words across three castes analysed in this study:

Table 3.4 Some examples of Marwari words as they appear in different varieties.

Variety	To say	I	my	What	water	where	road
Bishnoi	gʰɛɳo	mʰʊ	mero	kja	paɳi:	kɪʈʰ	həɖək
Brahmin	keɳo	mʰɛ	mʰɔro	kjc	pɔ̃ɳi	kəʈʰe	səɖək
Jaat	keɳo	mʰʊ	mʰaro	kja	paɳi:	kəʈʰe	səɖək

These caste groups have not been previously studied, so the data is based on researcher's native expertise and fieldwork with speakers, controlling for geography and gender. However, research on other languages suggests systematic phonetic and phonological differences frequently occur across Brahmin versus non-Brahmin varieties (Ruback and Rao, 1989; Chavan, 2013). A recent acoustic study also found vowel quality contrasts between Marwari-speaking Brahmins and Bishnois (Suthar, 2018).

Other than phonetic differences, many morphemic and vocabulary differences were observed during fieldwork between different castes. For example, the verb “to eat” has different words in these three varieties, i.e., /ji:mɳo/ for Bishnoi and Brahmin and /kʰaɳo/ for Jaat.

3.6 Rationale behind caste selection

The rationale for selecting these castes was twofold. Firstly, there has been no prior research on caste-based acoustic differences in Marwari spoken in Bikaner. These specific castes were chosen purely due to accessibility constraints and availability of speakers willing to participate. Secondly, Bikaner’s intricate caste origins means many have intertwined histories - for example, shared roots between Rajput and Charan or between Jaat and Bishnoi. By including such historically-linked castes, the researcher aimed to examine if acoustic measures demonstrate persistent similarities alongside variation across the varna hierarchy. Specifically, it was predicted that phonemic inventories would align closely across the two groups, including comparable individual vowel spaces. However, while originating caste relationships informed selections, the sampling was opportunistic based solely on researcher access rather than controlled demographic design. Any observed patterns may guide future controlled studies rather than make definitive community generalisations.

The data shown in the previous section to show phonological differences in the caste is based on the author’s personal native speaker expertise and is controlled for geography, caste, and gender.

The present study will focus on the acoustic differences between the selected vowels for each of these castes. The selection of vowels was based on their availability across castes. It should be noted that no phonological or morphosyntactic analysis is undertaken here. Additionally, while the researcher’s own caste identity may have shaped social networks and participant accessibility, by exclusion of the researcher’s caste variety limited any linguistic influence.

The goal is examining acoustic variation for inter-speaker differences across available vowels without aiming to make broader linguistic generalisation. By controlling factors such as region, gender and utilising native speaker’s intuitions, observed patterns highlight potential caste-based distinctions worth further investigation for the study.

The next chapter details fieldwork data collection methods.

4 Data Collection

The current study aims to evaluate the effects of incorporating within-formant vowel features in speaker comparison, utilising the Marwari language as a testbed. This chapter describes how the data collection for the study was carried out.

Section 4.1 provides information on the participant population and the rationale behind the selection of this population. Section 4.2 explains the ethical considerations made for the data collection process. Section 4.3 provides an account of the material employed in the fieldwork, encompassing the various data collection methods. The next section (4.4) describes the specifications of the recording equipment used during the fieldwork.

4.1 Participants

Speakers from three different caste varieties were selected (see Chapter 3 for additional details on caste). While caste differences themselves are salient, other social factors linked to these varieties cannot be discounted. A speaker's idiolect or personal style can be shaped by various other factors including audience/interlocutor accommodation (e.g. Bell, 1984; Pardo et al., 2022), race (e.g. Holliday & Squires, 2020), gender (e.g. Kiesling, 2002), and identification with external social groups (e.g. Labov, 1973). Since Labov's pioneering work on the social stratification of English, linguists have had a valuable framework for investigating linguistic variables such as social class, age, gender, and more (Labov, 2006). The present research was designed while accounting for both social (caste, gender, region, and age) and stylistic (different speech styles) variation. Sections 4.1.1 to 4.1.4 will elaborate on these selection criteria. The selection was based on region (Section 4.1.1), gender (Section 4.1.2), age (Section 4.1.3) and education (Section 4.1.4).

For Marwari, this distinction is based on "caste" rather than class. Chapter 3 (section 3.5) explicates caste as a social construct that evolved historically, chiefly shaped by occupational roles. Moreover, the survival and persistence of castes can be attributed to individuals of the same caste aligning to maintain distinct linguistic and cultural variations of their caste. Rather like social class, caste affiliations interact with other factors such as age, gender and region in determining or influencing language variation (Meena, 2015). In light of this, sections 4.1.1 - 4.1.4 provide an overview of these variables as they relate to participant selection.

4.1.1 Region and dialect

Geographical region is an important predictor of dialectal diversity (Moosmiller, 1997). However, in the latter half of the twentieth century, there has been a debate surrounding the persistence of regional dialects in the face of constant migration (Chambers, 1994). The contemporary interpretation of the concept of region has evolved from being the 'most important' factor in determining a dialect to being acknowledged as 'one of the possible factors' in shaping a dialect (Chambers, 2000; Moosmiiller, 1997).

Based on availability and access, three distinct castes were each divided into groups of fifteen people.

Speakers from the Bishnoi and Jaat varieties predominantly reside in the rural areas of the Bikaner district. Brahmins, on the other hand, live in urban areas. The majority of the participants were homemakers who seldom left their neighbourhoods or interacted with individuals outside their communities. Section 4.4. will offer a detailed review of the recording circumstances.

The next section provides a comprehensive overview of the gender-specific information of the participants.

4.1.2 Gender

The decision to exclusively recruit female participants in this study is rooted in a three-fold rationale. Firstly, as elaborated in section 4.1.1, the core objective of the study is to discern speaker-specific features based on vowel formants. Given that the data collection could not occur in controlled laboratory settings, additional measures were introduced to ensure data consistency, such as participant monolingualism. By opting for female participants, the study sought to minimise any potential influence from second languages or dialects in a multilingual nation such as India. The adult female population within the speech community met this criterion to a considerable extent.

It should be noted that a speaker's self-reported linguistic knowledge or monolingualism was determined by a questionnaire filled out during field work. Despite this method, the possibility that speakers comprehend more than one language cannot be denied, especially in a

multilingual country such as India. However, as supported by their questionnaire for the purpose of this study they will be treated as monolinguals.

Secondly, practical reasons, including convenience, shaped the focus on females. Notably, male participants were either absent or had scheduling conflicts during the data-collecting period. Moreover, due to socio-cultural reasons, males are more influenced by additional languages, such as Hindi, rendering them less appropriate for the present investigation. Theoretically, the same research could be conducted with multilingual male speakers by developing an appropriate model.

The third rationale for female participants stems from the underrepresentation of female speech in acoustic analyses for speaker comparison. While the role of vowel formants in gender identification has been explored (Labov, 1973), with male formants typically being lower than female formants (Bachorowski & Owren, 1999), studies of female speech are conspicuous in their absence in respect of speaker comparison research. This research seeks to redress this gap through analysis of within-formant features for previously under-researched female speech by focusing only on women.

The next section will provide a detailed overview of the age-related information of the participants.

4.1.3 Age

The study's participant selection was guided by their age group, specifically targeting individuals aged 40 and above. This criterion was established with the hypothesis that this group's language would have minimal external influences (as the monolingual female population is largely confined to its home community). It is important to note that the selection process only considered age as one of the criteria and once the feature analysis started the study did not delve into a detailed analysis of the data based on various age subgroups within this selected category. However, any conclusions are assumed to be influenced by age-related factors throughout time.

The average age of the participants in this research was 50.68, ranging from 40 to 84, with a standard variation of 8.03. Participants varied in age from 40-50 for Bishnoi, 40-65 for Brahmin, and 45-84 for Jaat. The majority of participants in all three varieties were between the ages of 40 and 70, with just one exception.

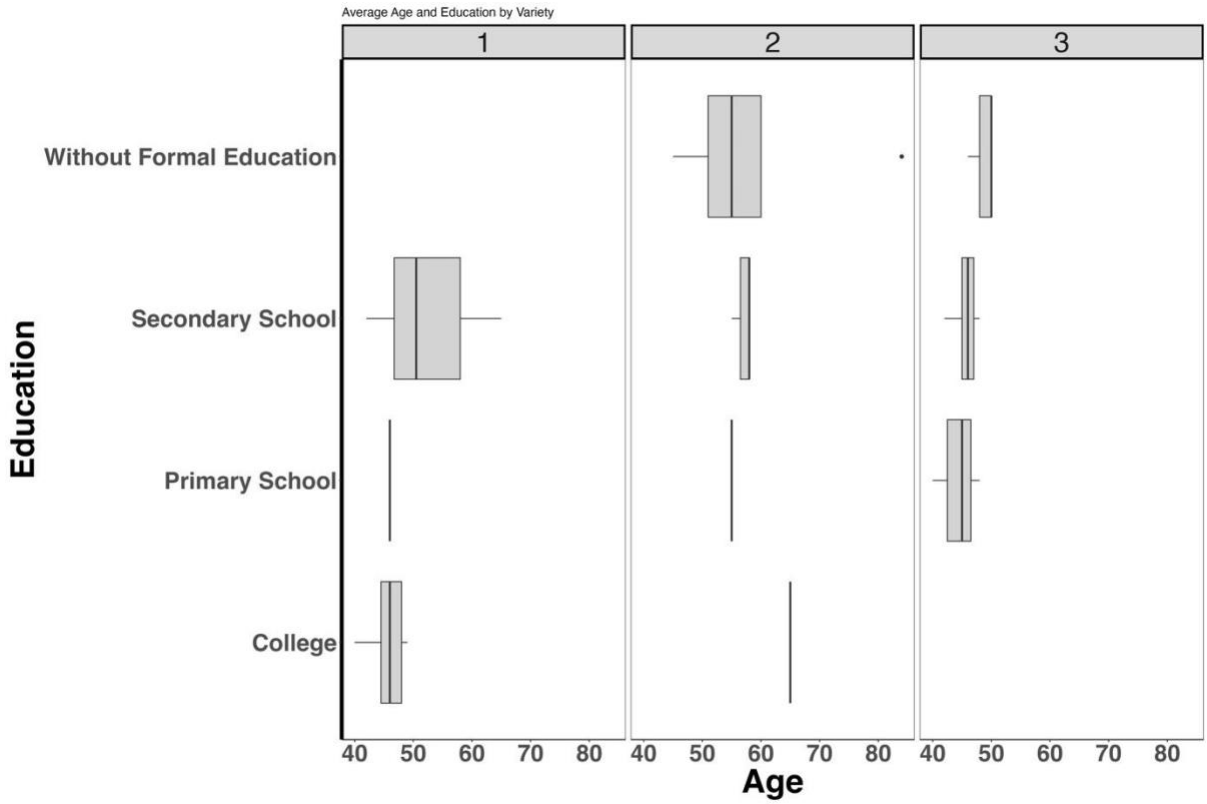


Figure 4.1 Average age and educational background of the participants (1= Brahmin variety, 2 = Jaat variety and 3 = Bishnoi Variety)

The next section will provide the educational background of the participants.

4.1.4 Education

Participants in the study were surveyed regarding their educational background and range of languages. The inquiry into their educational backgrounds held significance because, as elucidated in Chapter 3, the primary mode of education in rural Bikaner is Hindi. However, teachers often use Marwari as a teaching language to facilitate better communication with their students. This practice directly influences the linguistic competence of the participants, as they acquire proficiency in either Hindi or the standard variety of Marwari during their school years.

Participants were recruited under the assumption that individuals in the area would possess familiarity with their respective caste dialects. These selection criteria were contingent upon their knowledge of the caste dialect, effectively limiting participation to monolingual individuals during the data collection process. Despite the rigorous efforts to recruit monolingual participants, finding individuals who met this criterion, especially within the

Brahmin variety, posed challenges. Notably, the majority of Brahmin participants had finished at least a high school education (10th grade). The study recognises the possible effect of schooling and will take it into account when examining caste- or variety-specific differences.

As presented in Figure 4.1, educational qualification was divided into without formal education (participants who never went to school), primary (participants who had some kind of formal primary education, i.e., up to 5th grade), secondary (participants who finished high school), and college (participants who went to college). During the fieldwork, the highest qualification documented was completion of an undergraduate degree. The reading skill of participants was determined based on their educational backgrounds. Anyone who attended any type of official educational system, even primary school, could read, and hence could complete their wordlist task without the assistance of an informant (please refer to section 4.3 for further details on these tasks)

For Brahmin variety, none of the participants were without any education (everyone could read), one participant had at least some primary school education, 6 participants attended secondary school, and around 8 participants attended college.

For Jaat variety, 10 participants had no formal education, one went to elementary school and 3 went to secondary school, and one went to college.

For Bishnoi variety, 7 participants had no formal education, 3 participants attended elementary school, 5 participants attended secondary school, and none attended college.

Overall, 17 out of 45 individuals were unable to read; 14 had a secondary education, 9 had a college education, and 5 had a primary education.

As previously discussed in section 4.1.1, the first criterion for participant selection was their monolingualism. Each participant recruited for the study affirmed their exclusive ability to communicate solely in their respective caste dialect. Based on participant information collected, none of the participants reported fluency in either the standard dialect of Marwari or Hindi, encompassing both spoken and written forms. Conversely, the Brahmin variety exhibits a greater concentration of speakers in the upper half of the educational spectrum. It's worth mentioning that none of the Bishnoi participants claimed to have any university or college education.

As a Marwari speaker with links to the Bikaner district, the researcher had some connections with the informants of these communities, and with their assistance, suitable participants were chosen.

The methodologies employed for data collection, particularly with regard to participants who may be considered naïve or less familiar with research procedures, were subject to rigorous ethical clearance protocols. The design of fieldwork was crafted with paramount consideration for the safety and dignity of the participants, and section 4.2 will discuss these ethical issues in further detail.

4.2 Ethical Approvals

Before data collection, each participant completed a written consent form (see Appendix 10.1 for an example of the consent form) to ensure that the research was respectful and ethical for both participants and researchers. The permission form was in the Devanagari script. The Devnagari script is the traditional writing system used by Indo-Aryan speakers in the region.

This form was presented to and approved by the University of York ethics committee prior to the commencement of fieldwork. The researcher and a family member read the forms aloud to the participants who could not read them and helped them comprehend the goal of the study. Because the majority of them could not write, they signed the permission form with a thumbprint. All participants were unaware of the narrowly defined research topic but were aware of the overall goal of the study. The research adheres to the guidelines provided by the University of York.

The fieldwork began on December 13th, 2019, and finished on December 30th, 2019. The data was gathered from 45 people, 15 from each of the three castes. During the fieldwork, it was discovered that the majority of the participants were quite shy but eager to submit data. Further, they were very pleased to be a part of this experiment.

4.3 Materials

The recordings include both spontaneous and non-spontaneous speech. The initial technique of data collection was a wordlist, in which was designed to be read aloud from the Devanagari script (a type of written script used by Marwari speakers). Non-readers were assigned an informant

from their own variety to read these words aloud for them, which were subsequently repeated by the participants. (For detailed information, please see section 4.3.1.)

The second mode was a picture description task, in which participants were shown an image of a local god and asked to provide a tale about the deity. The third task was a natural conversation, in which two participants were paired and invited to engage in unscripted talk on a topic of their choosing or one picked from a list supplied. Some of the issues discussed were - childhood, marriage, daily routine, shopping, farming, and family.

The word list task was completed by participants in around two minutes. The story task was allotted 10 minutes, but most participants finished in two to three minutes. The conversation task was given 10 minutes for two participants to converse together (aiming for around five minutes of conversation data per participant). The recordings ended after 10 minutes of conversation time. Sections 4.3.1 through 4.3.3 will discuss these tasks in further detail.

4.3.1 Wordlist

When conducting fieldwork for acoustic analysis, it is important to understand how to obtain recordings and what to record. This study encompassed both spontaneous and non-spontaneous speech sounds, with a specific focus on vowels spoken by all three caste varieties. To retrieve a non-spontaneous recording, a wordlist was created.

Due to a lack of prior work or prepared wordlists for phonetic and acoustic analysis of this language, a new wordlist was devised. Swadesh (1955) emphasised the importance of carefully selecting wordlist meanings for accurate lexico-statistic dating. He proposed a 200-word list suitable for comparing languages and estimating divergence dates. For this study, Swadesh's wordlist served as a starting point for creating a Marwari version. Words were chosen from his list and translated into equivalent terms in Marwari. This provided a customised wordlist in Marwari that drew upon Swadesh's research on optimising wordlists for language comparison and dating.

The next step involved securing words with vowels in the required position, i.e., CVC (vowel as the syllable nucleus within single-syllable words, with consonants proceeding and following the vowels). This approach aligned with the methodologies employed by earlier researchers (Labov, 1973; Ladefoged, 2003). Adi-Bensaid and Tobin (2010) recommended placing the target vowel between two obstruents to prevent any vowel lengthening, ensuring a clear

beginning and ending for every vowel. This process effectively mitigated the observed vowel lengthening that often occurs in isolated words (Hildebrandt, 2005). Importantly, only culturally appropriate, and locally relevant words were selected for inclusion in the wordlist.

According to the 2011 Indian census, the female literacy rate for the rural population of the state of Rajasthan was reported as 52.12 per cent (Ram, 2014). However, with only 44.81 per cent of women possessing the ability to read or write their names in rural Bikaner, it was challenging to find participants who could effectively read the wordlist.

The initial task of reading the wordlist aloud was impacted by the fact that around 38 per cent of participants could not read the words. To address this issue, Ladefoged (2003)'s "community or group pronunciation" technique was used (pp.22-23). This technique is typically used when a researcher aims to identify variations within speakers by having a diverse group of individuals sit together, with a respected speaker uttering a word that the group then repeats. In the current context, instead of a respected adult speaker, a person from the same community who could read was chosen to articulate the word, which the participant would then repeat. Ideally, both individuals would pronounce the word identically, occasionally, the pronunciation of these words was subject to slight consonantal variations stemming from the idiolects of the assisting individuals. Notably, this phenomenon was particularly prevalent in the Bishnoi variety, where the informants, primarily individuals in their twenties, refrained from replacing the initial /s/ with /h/, a pattern commonly observed in Gujarati (Cardona & Suthar, 2014). This deviation from the usual Marwari pronunciation, wherein these consonants at word-initial positions are in complementary distribution, occasionally impacted the participants' pronunciations of the same words. However, at times, participants would express a strong desire to emphasise that their pronunciation of a word was correct, and any alternative pronunciation was deemed incorrect. It is noteworthy that these recordings were preserved in their original form, and any accents or pronunciations provided by the participants were retained to ensure the integrity of the data collection process.

A picture description task could also be used for such situations, but opting for a picture description task involving eighty target words would have also presented its own set of challenges. Firstly, data collection may have been hampered by the time-intensive process of identifying intended words based solely on pictures. Although the word list was derived from Swadesh (1955) and contained common daily terms, this issue would have been compounded

for Marwari caste varieties. Different castes often use divergent terminology, even for mundane objects. For example, Brahmins use /kəpɽɑː/ to denote clothing, while Jaats employ /puːr/ or /gɑːbbhɑː/ for garments. Relying solely on pictorial cues, without accounting for caste-specific lexicons, could have created confusion and introduced extraneous vocabulary. Streamlining the data collection process required preemptively addressing the potential complexities of caste dialects.

Additionally, with multiple words denoting the same object or concept, participants might have experienced experiment fatigue, leading to reduced attentiveness and potentially affecting the quality of data gathered. Given these potential issues, the decision to utilise alternative methods, such as the community pronunciation technique, was deemed more practical and efficient for the specific fieldwork context.

While repetition and mimicry of a standardized prompt provides a practical solution for collecting comparable wordlist data across castes, this approach risks obscuring fine-grained phonetic details and individual speech patterns. As Pardo et al. (2022) demonstrate, vocal accommodation is complex and variable. Forcing uniformity could distort vowels, which are very susceptible to convergence (Pellegrino & Dellwo, 2023; Kalmanovitch et al., 2015). It may also dampen subtle spectral qualities such as formant bandwidths (Pardo et al. 2013).

In summary, utilising a standardised prompt undoubtedly simplifies aggregation and analysis for broad lexicons. But for fine-grained phonetic study, especially of vowels, forcing identical repetition contradicts research showing idiolectic adaptation. While practical, mimicking a prompt does risk concealing the granular patterns, variations, and nuances that characterise Marwari caste dialects. As such, alternative approaches should be pursued if feasible to preserve the diversity of individual speech. Careful consideration of study goals and scope is warranted when weighing practical methodology versus ideal models for capturing phonetic intricacy.

4.3.2 Story

In the development of the fieldwork, the concept of Recorded Text Testing (RTT), originally devised by Casad (1987) for story data, served as a reference. According to Casad, RTT is a method for determining the inherent intelligibility of related dialects of a language by eliciting and recording a story in one dialect then evaluating speakers from another linguistic variety to

determine how well they understand it. However, a modification was implemented for this study. Instead of eliciting personal stories from participants, they were presented with various pictures and asked to craft a narrative associated with the depicted scene. As all of the participants were practising Hindus, they were able to recognise and were familiar with narratives concerning the deities.

To ensure the absence of any researcher-induced bias, participants were shown multiple images of deities. From this selection, they were free to choose one of the pictures and then narrate a story related to the deity's life. An intriguing observation was that many participants opted to narrate stories related to "Lord Ganesh" after viewing his image.

Figure 4.2 and Figure 4.3 provide visual representations of two such images employed during the fieldwork.

Figure 4.2 An image of God Ganesh shown to the participants for the story task (www.pixabay.com, 2023).

Figure 4.3 An image of Goddess Laxmi was shown to the participants for the story task (www.pixabay.com, 2023).

4.3.3 Conversation

The third task of the fieldwork involved collecting spontaneous speech by recording conversations between two participants. To minimise the impact of the observer's paradox[6] (Labov, 1972), this task was strategically scheduled for the conclusion of the fieldwork. By this point, participants had already completed the other two tasks, becoming more familiar with the researcher and the recording equipment. Consequently, they were generally calmer and less self-conscious during these recorded conversations.

A list of conversation topics had been prepared by the researcher in case participants were unsure about what to discuss with each other. However, interestingly, none of the participants requested or relied on the prepared list. Figure 4.4 illustrates a participant from the Bishnoi community, and Figure 4.5 shows the participants from the Brahmin community having a conversation with each other.

[6] Labov (1972) outlines an observer's paradox as a phenomenon in which people change their speech patterns when they are aware that they are being observed. This makes it difficult for linguists to collect authentic, unfiltered data on how individuals converse in ordinary situations because their knowledge of being observed might produce speech changes.

To reduce confusion, the informant left the room after assisting the participant with the word list task, since they were only needed for that portion of the study.

Figure 4.4 A picture of Bishnoi participant (printed with permission of the subject).

Figure 4.5 A picture of two Brahmin participants having a conversation (printed with permission of the subjects).

To prevent the participants from becoming self-conscious or altering their behaviour, these images were obtained after the participants had completed their assigned tasks. This approach ensured that their responses and interactions during the tasks remained as natural and

unaffected as possible. The decision to install recording equipment in each household was carefully considered, with the awareness that such technology could potentially make participants feel awkward and overly aware of their surroundings, thereby influencing the data to maintain a conducive and distraction-free environment for participants, the avoidance of any additional external technology, including photography, was prioritised during the fieldwork. Section 4. will provide a comprehensive understanding of the recording settings and equipment employed during the data collection process.

4.4 Recordings

Participants were recorded under controlled conditions, utilising a high-quality digital recording device, 'Zoom H4n Handy Recorder'. The initial recordings were collected in wav. format, with 44.1 kHz sampling rate and 16-bit depth and were all stereo. This recorder was equipped with built-in microphones that could be adjusted to either a 90-degree or 120-degree configuration.

During the recordings, two different channels on the recorder were used for two reasons. First, this setup allowed for the simultaneous collection of samples from individual participants for the third part of the data collection, i.e., conversational recordings, when two participants were engaged in dialogue – one was recorded on the right channel of the sound file, the other on the left. Second, it ensured that there was a suitable selection in at least one channel for every recording, enhancing data quality and consistency.

The recorder was positioned at a standardised distance of 25 centimetres from the participant's mouth, secured on a tripod to maintain uniform recording conditions for every participant. The microphone settings on the recorder were adjusted to a 120-degrees configuration for both channels, depending on the participant's position. These controlled conditions were characterised by a quiet environment within the participant's house, minimising distraction and noise to facilitate high-quality recordings.

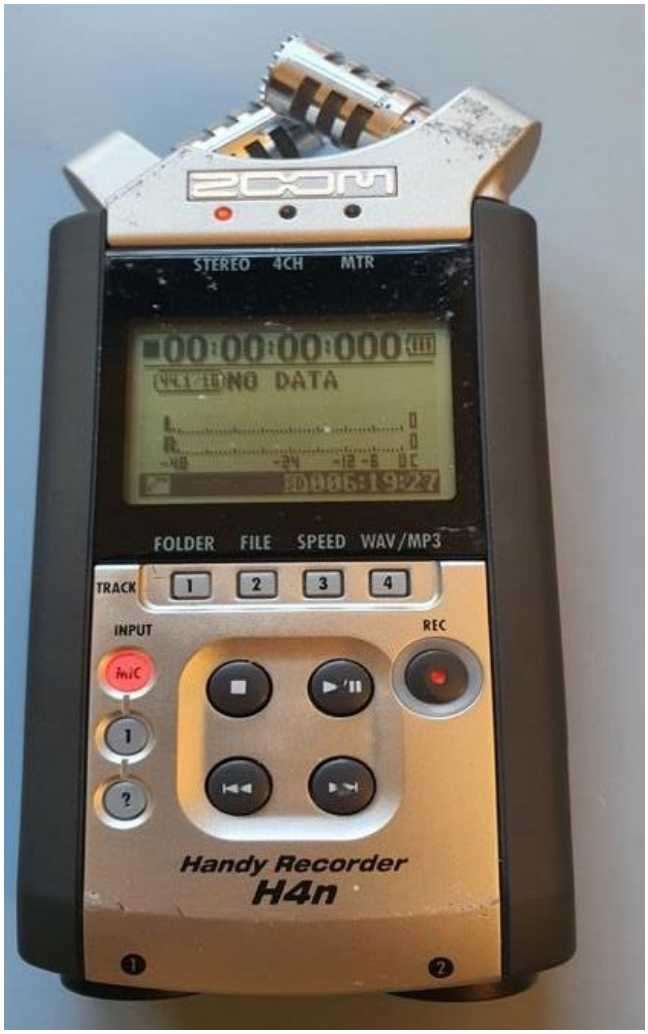


Figure 4.6 An image of the recorder used during the fieldwork.

In the research process, eight vowels were chosen for each variety. The selected vowels were: /a:/, /i:/, /u:/, /ʊ/, /o/, /e/, /ə/, and /ɪ / (see chapter 3 for further details).

For the subsequent step, ten tokens of each of these vowels were selected by listening to the recordings and using native speaker intuitions. These tokens were then combined into a new WAV. file. This combining process was facilitated using the software Soundforge (9.0).

Each vowel was situated within a CVC (consonant-vowel-consonant) syllabic structure (see appendix table for wordlist). Moreover, both the coda and onset positions of the selected syllable included a range of consonants, both voiced and voiceless and with varying places and manners of articulation,

The audio recordings were made over the duration of one month from a group of 45 different speakers. To maintain consistency and minimise external influences, all three modes of data collection for each individual were conducted in a single session. This approach aimed to mitigate potential sources of interference, including emotional, biological, or physical stress.

The original intention was to gather two sets of recordings for all three data collection modes separated with a one-year interval between them. However, this plan had to be abandoned due to the onset of the COVID-19 pandemic, which occurred just three months after the initial recordings were obtained. It was therefore impossible to include non-contemporaneous recordings of each participant.

In the context of fieldwork in India, mitigating background noise presents a substantial problem for researchers. Following the recommendations of Chelliah and Reuse (2011), participants were requested to turn off ceiling fans and coolers while keeping windows closed during the data collection sessions. However, it is worth noting that for Brahmin participants, additional sources of noise, such as ongoing construction work and vegetable vendors, occurred even though efforts were made to minimise these disruptions.

In contrast, for Bishnoi participants, who primarily resided in quieter farming communities, these types of disturbances were not a concern. The selection of recording rooms considered their distance from the street and the amount of furniture within them, with the aim of minimising any potential echo, as recommended by Ladefoged (2003).

4.5 Summary and Discussion

The methods employed for the data collection yielded three types of speech data for analysis from each of the three caste dialects under study.

Unavoidable imperfections in the data elicitation and recording processes determined by the domestic, social and educational backgrounds of participants (e.g., extraneous noise, differing acoustic environments, non-literacy) are acknowledged and have been borne in mind in analysing and interpreting the data in the chapters following.

A second round of fieldwork – originally planned but prevented by the lockdown and travel restrictions associated with the COVID-19 pandemic – would have provided non-contemporaneous speech from the participants and therefore enabled an assessment of the

degree to which intra-individual speech measures were stable. With the agreement of my advisors and my funding body, I have taken steps to compensate for this by subjecting the ‘one field visit’ data to more intensive and comprehensive analysis than would have been possible had things gone to plan.

C’est la vie!

5 Data Processing

This chapter is segmented into four sections. The first section provides a brief account of the vowel extraction process and pre-processing steps employed for all three modes of data elicitation. The second section examines initial vowel analysis across varieties to ascertain variety-specific differences. The third section discusses the within-formant features extracted from vowel formant mid-points for each participant. The final section examines if vowels, varieties or modes of data elicitation significantly impacted feature values.

5.1 Isolating Target Sounds

Data processing commenced by isolating the target sound files per participant. This comprised three stages: identifying required sound sections (words) from recordings, processing to remove sections with background noise using Sound Forge (9.0) and analysing with Praat.

At the first stage, words with required vowels were isolated by removing unwanted sounds into a new sound file, then further checked and sections with background noise were discarded. Finally, the clearest channel from the recording was extracted with Soundforge. The recordings were collected in 'stereo' mode, which were then converted to 'mono' at this stage. The next step was to identify the clearest channel with the least amount of background noise.

Peaks were gain normalised in Soundforge to 2.0 dBFS (decibels relative to full scale) to bring the overall loudness to a certain level without clipping the sound (Jessen, 2008). Following this, the normalised channel was saved on a new '*.wav'* file for analysis.

5.1.1 Vowel extraction

The next step was extracting selected vowels from sound files using Praat (6.1.54) (Boersma & Weenink, 2001) and a script developed by Dr. Philip Harrison (Harrison, 2019). Vowel extraction from CVC syllables with obstruents at both coda and onset locations was used during segmentation (Adi-Bensaid & Tobin, 2010). Their work proposed that using obstruents over sonorants reduces vowel shortening and preserves quality. Syllable boundaries were marked using both waveforms and spectrograms. The settings included in the script changed the default spectrum and formant settings provided in Praat to ones provided in the script. Additional manual changes were made for better visualisation. These adjustments included changing the pre-emphasis, dynamic range, and maximum spectrum view settings. Figure 5.1 illustrates the

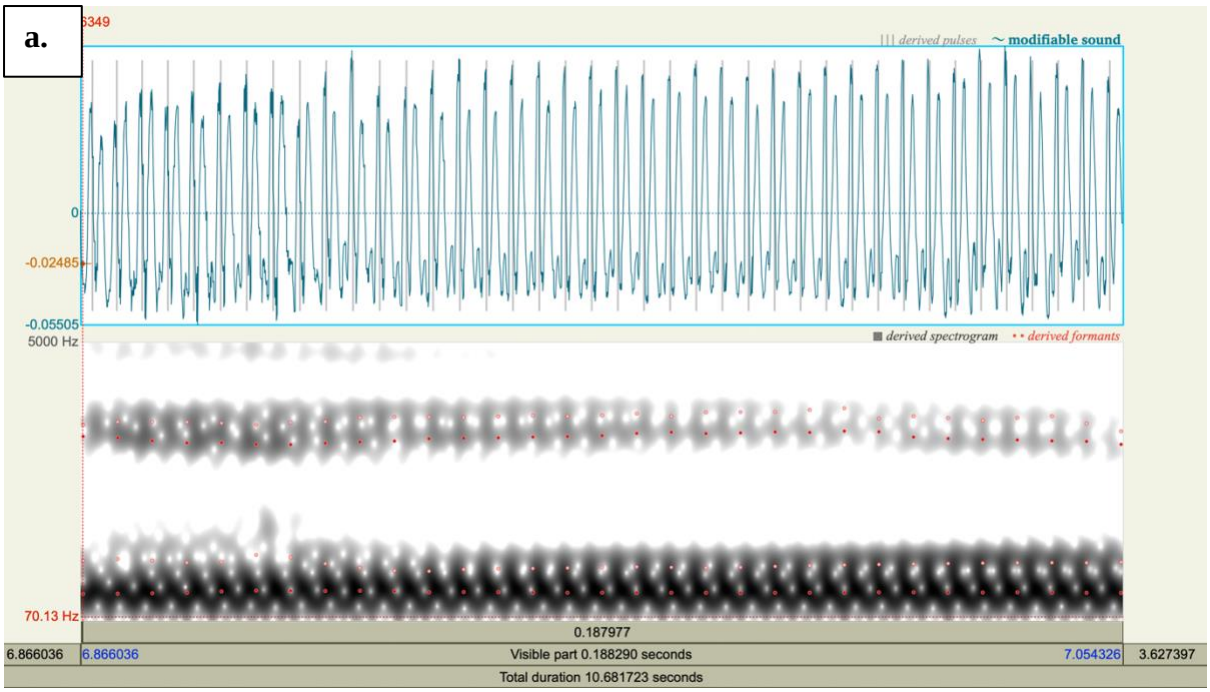
impact of these modifications, as altering the dynamic range from 30 (a) to 60 (b) made the formant energy levels more clearly visible.

The final settings were -- window shape - ‘Gaussian’, maximum spectrum view - ‘100 Hz’, pre-emphasis - 6.0 dB, and method of spectrum analysis - ‘Fourier’. These settings did not affect the audio signal and were only used to get an optimum view of the spectrum. For formant analysis settings, the formant ceiling was set to 5000 Hz (i.e., the maximum number of visible formants in a spectrogram would be up to 5000 Hz).

The dynamic range was set to 30 dB. This script logged individual formant frequencies up to F4 per vowel, and the differences between them. Values were saved as .tab files, which were easily accessible in Excel or any spreadsheet format. Logged measurements included extraction details such as pre-emphasis uses for future reference (if needed).

Individual log files for each vowel were also provided by the script, which included details such as formant averages, standard deviations, range, and minimum, and maximum values.

Figure 5.2 shows an example of the formant extraction procedure. The formants were retrieved from two pulses and averaged from vowel midpoint. There were some cases where the formant tracking provided by Praat’s inbuilt formant measuring tool did not align with the formants in the spectrograms, and for any such analysis, the formants were manually logged based on visual observation Figure 5.3 shows an example of the output text file. Figure 5.4 provides one example case. In the figure, we can see two versions of the same sound. In both Praat is picking up all four formants, but the formants are not falling on their respective positions in the spectrogram. The version (a) shows the spectrum without formant tracks and version (b) marks the formant tracks. This is because the script is unable to detect the formants automatically. In such cases, some manual measurements based on visual observations were extracted and inserted for the formant data.



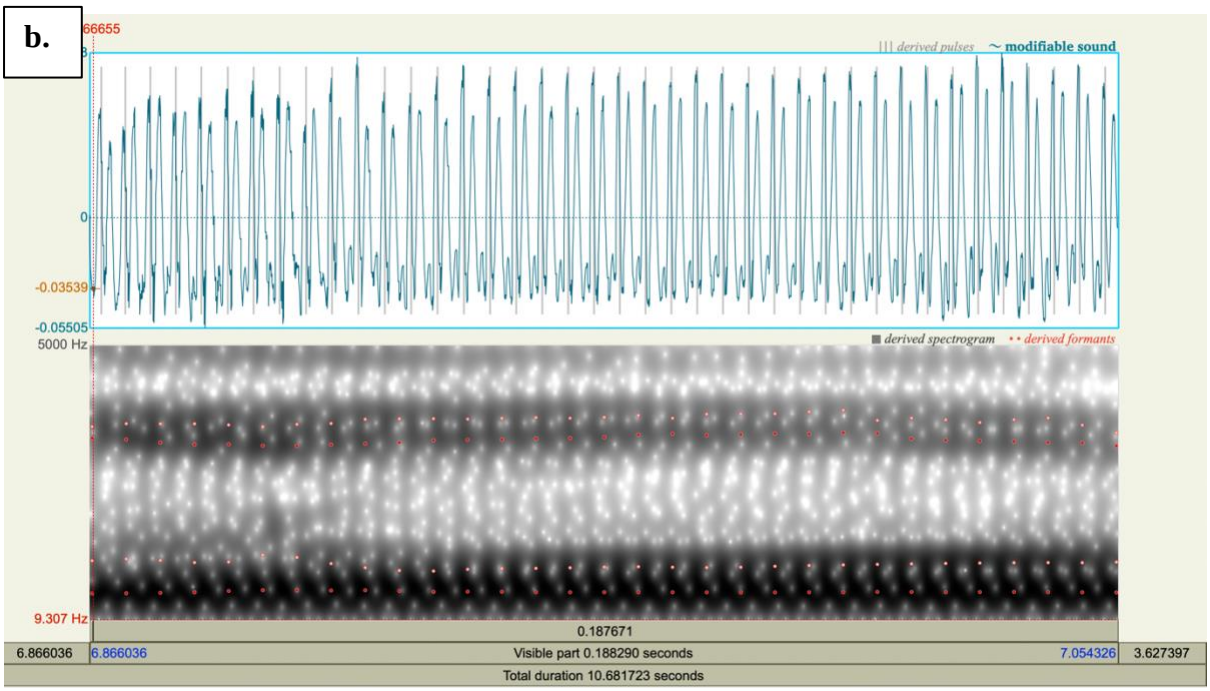


Figure 5.1 (a.) An example of sound /o/ with pre-emphasis set to 3 dB, **dynamic range 30 dB** and maximum spectrum view 100 dB. (b.) An example of sound /o/ with pre-emphasis set to 3 dB, **dynamic range 60 dB** and maximum spectrum view 100 dB.

Table 5.1 Initial formant measurement used for formant extraction.

Maximum number of formants	4.5
Maximum formant	5000
Dynamic range	30
Window length	0.025

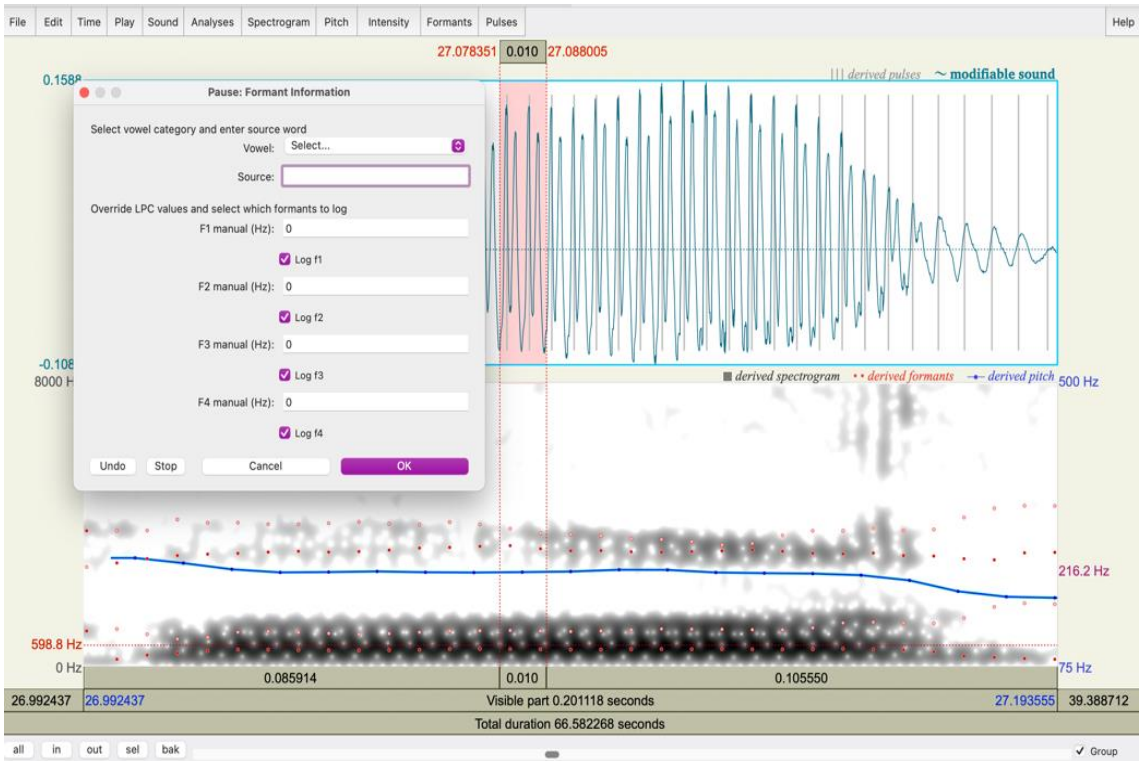


Figure 5.2 An example image of the extraction process using Praat script (Boersma & Weenink, 2001).

	1	2	3	4	5	6	7	8	9	10	11	12	13
row	f1	f2	f3	f4	f4-f3	f3-f2	f3-f1	f2-f1	From	To	Vowel	Source	Max_Fo
1	388.0711366	2536.236251	3271.568915	4433.798285	?	735.3326634	2883.497778	2148.165115	4.567697743	4.577426702	i:	bibali	500
2	378.3798999	2557.936448	3187.754589	4294.974952	?	629.8181411	2809.374689	2179.556548	5.621116134	5.634230685	i:	diid	500
3	374.1913349	2614.531351	3128.374314	4566.329411	?	513.8429632	2754.182979	2240.340016	6.299816347	6.311088262	i:	gigaliya	500
4	679.1910816	2224.957666	3229.685044	4541.0376	?	1004.727379	2550.493963	1545.766584	7.288455307	7.300051875	I	chinch	500
5	356.5125914	2336.239922	2933.603796	4561.14713	?	597.363874	2577.091205	1979.727331	7.932207632	7.944975689	i:	jhijhak	500
6	491.024178	1828.703606	2837.948216	4582.053319	?	1009.244609	2346.924038	1337.679428	8.132243153	8.143353547	c	jhijhak	500
7	425.2425258	2445.777539	2880.792529	4074.083167	?	435.0149903	2455.550003	2020.535013	8.610925438	8.622411873	I	mimli	500
8	370.3224408	2494.415119	3002.598436	4625.534467	?	508.1833169	2632.275996	2124.092679	9.948678305	9.965191345	i:	riri	500
9	386.9184805	2582.033407	3163.045422	4524.980945	?	581.0120149	2776.126942	2195.114927	10.15520352	10.16963477	i:	riri	500
10	382.5436174	2363.1084	2971.200523	4581.157144	?	608.0921236	2588.656906	1980.564782	10.49335775	10.50530278	i:	shisho	500
11	438.1581465	2381.707306	2925.553401	3857.771316	?	543.8460942	2487.395254	1943.54916	11.23561558	11.24728473	e	dedi	500
12	337.6199934	2667.072844	3379.253442	4560.467825	?	712.1805971	3041.633448	2329.452851	11.45406485	11.46636513	i:	dedi	500

Figure 5.3 An example table containing formant values from F1-F4 with the difference between each formant and the extraction information logged in by the Praat script used.

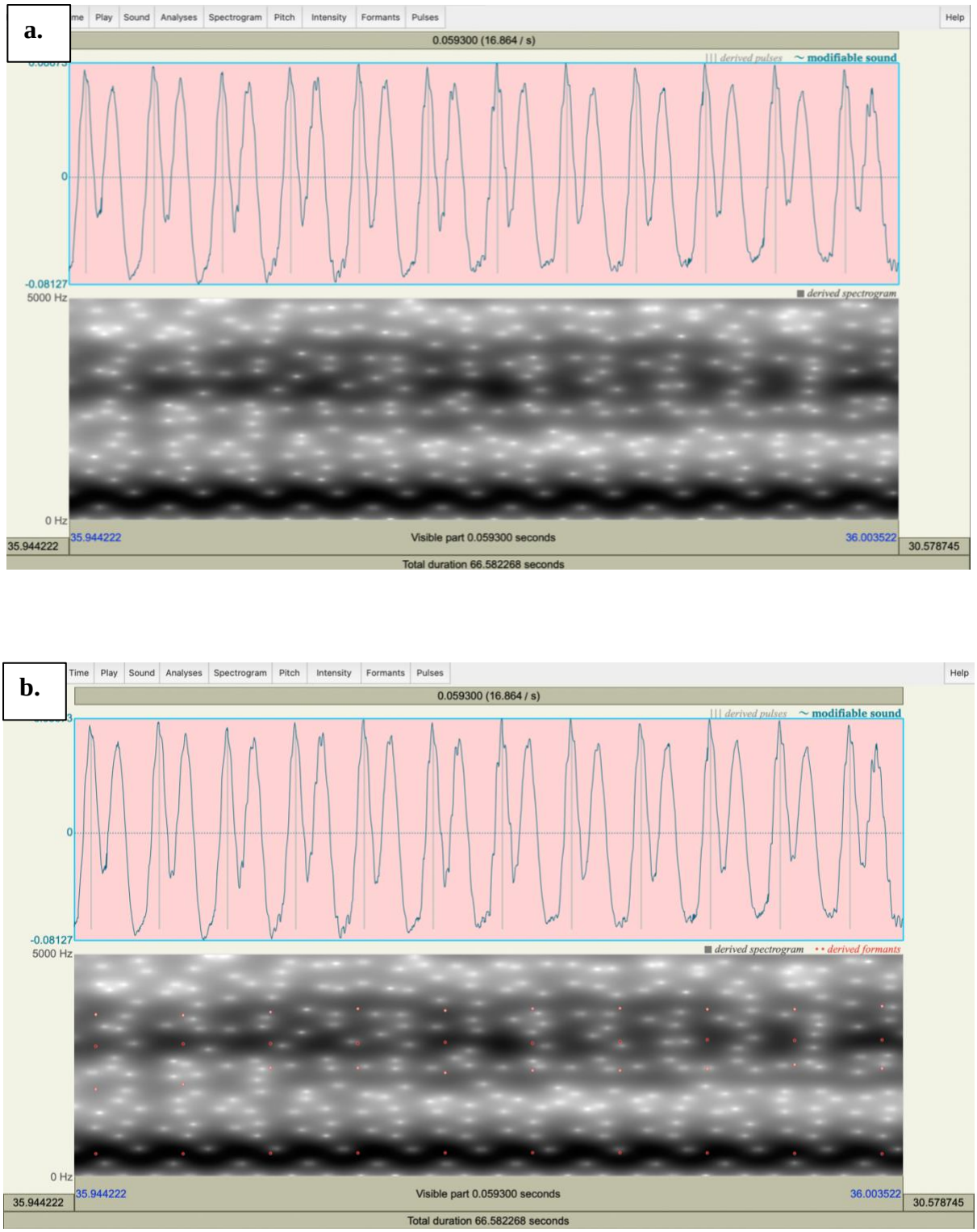


Figure 5.4 A spectrogram of sound /e/ without formant tracks (a.) and with formant tracks where the formant markers for F2 is not aligned with its respective formant.

Ten tokens were extracted for each vowel (for each mode of data elicitation), resulting in 80 tokens per participant and 1200 tokens for each vowel in a single variety. Table 5.2 offers a

comprehensive overview of each variety along with the tokens produced by specific vowels. On average, for each variety, a minimum of 97 per cent of tokens were considered usable. Notably, back vowels had the highest number of discarded measurements, as illustrated in the table, with vowels /u:/ and /o/ from Jaat variety having the most discarded tokens (18 and 12, respectively). This was likely due to the close proximity of F1 and F2, causing them to be interpreted as a single formant. both manually and automatically. The percentage indicated reflects the proportion of tokens deemed unstable after the completion of both manual and automatic extraction processes. Additionally, some tokens were excluded because the boundaries of formants were not clearly discernible.

Table 5.2 Number of tokens extracted for individual varieties and vowels.

Variety	Total tokens	/a:/	/ə/	/e/	/ɪ/	/i:/	/o/	/u/	/u:/
Brahmin	1181(98%)	150	150	150	145	145	150	146	143
Jaat	1148(96%)	143	145	150	145	150	138	145	132
Bishnoi	1193(99%)	150	145	148	150	150	150	148	148

5.2 Initial Analyses

5.2.1 Vowel space chart for individual varieties

Visual representations in the form of vowel space charts of each variety were created. These charts facilitated the comparison of different varieties.

To generate a vowel space chart, the initial step was to clean and normalise the data. This process was carried out in R (R core team, 2023). The F1 and F2 values were normalised using the z-score normalisation method. Z-score normalisation (or standardisation) is a statistical approach that rescales a distribution with a mean of zero and a standard deviation of one (Diez, Cetinkaya & Dorazio, 2015). This technique transformed each element in a dataset into a typical normal distribution. To get the z-score of a data point 'x' in a distribution with mean 'μ' and standard deviation σ, following formula was utilised:

$$Z = \frac{x - \mu}{\sigma}$$ (Diez, Cetinkaya & Dorazio, 2015)

In R, to normalise the data with z-scores, the 'scale' and 'abs' functions were applied to each column containing formant values. Firstly, the 'scales' function scaled the formant values to a mean of 0 and a standard deviation of 1, and later the 'abs()' function converted these scaled values to non-negative absolute integer. The outcomes were assigned to a new column in the data frame, which represented the original column's z-score normalised values. Finally, any possible outliers were eliminated by eliminating the items in the new formant columns with the corresponding values larger than 3.29 (indicating prospective outliers).

The next step was to retrieve variety-specific averages for each vowel and visualise them using vowel space charts. Table 5.3 shows averaged first and second formant (F1 and F2) frequencies in Hz for eight vowels (/ɪ/, /i:/, /e/, /ə/, /ɑ:/, /o/, /ʊ/, /u:/) produced by Brahmin, Jaat, and Bishnoi speakers. The average values were calculated for all data types together here.

Overall, F1 values correspond closely across groups indicating similar vowel height distinctions. However, F2 patterns show systematic frequency shifts reflecting front-back distinctions which differ among the three groups. For high front vowels /ɪ/ and /i:/, Bishnois use a more backed variant than Brahmins and Jaats whose realizations are similar. In mid and back vowels (/e/, /o/, /ʊ/,/u:/), the Brahmin group consistently produces more fronted variants while Bishnois use more backed realizations. Mid central vowel /ə/ also patterns in the same way. For low vowel /a:/, Jaats have a more fronted articulation versus more backed variants from Brahmins and Bishnois. These formant frequency shifts reveal subtle dialectal variations in tongue body positioning during articulation for the same vowel targets among these three language sub-groups.

Table 5.3 Variety-based averages of F1 and F2 values for extracted vowels.

Vowel	Brahmin				Jaat				Bishnoi			
	F2	F1	F2SD	F1SD	F2	F1	F2SD	F1SD	F2	F1	F2SD	F1SD
ɪ	2334	420	305	78	2385	430	299	76	2281	417	312	8
i:	2472	430	260	73	2516	431	221	71	2441	419	274	74
e	2193	549	263	97	2241	527	290	98	2136	530	275	102
ə	1701	577	262	109	1717	574	244	102	1673	545	269	121
ɑ:	1566	727	203	99	1647	755	190	126	1562	704	225	121
o	1225	520	230	89	1195	515	237	83	1211	510	237	83
ʊ	1205	445	195	87	1166	438	189	80	1195	444	187	83
u:	1094	442	167	90	1028	436	172	81	1084	443	189	94

To move further into vowel characteristics, F2 versus F1 plots were generated as vowel space charts presenting differences between the vowels of the three varieties. This step helps visualise the average vowel space occupied by an average speaker from each variety.

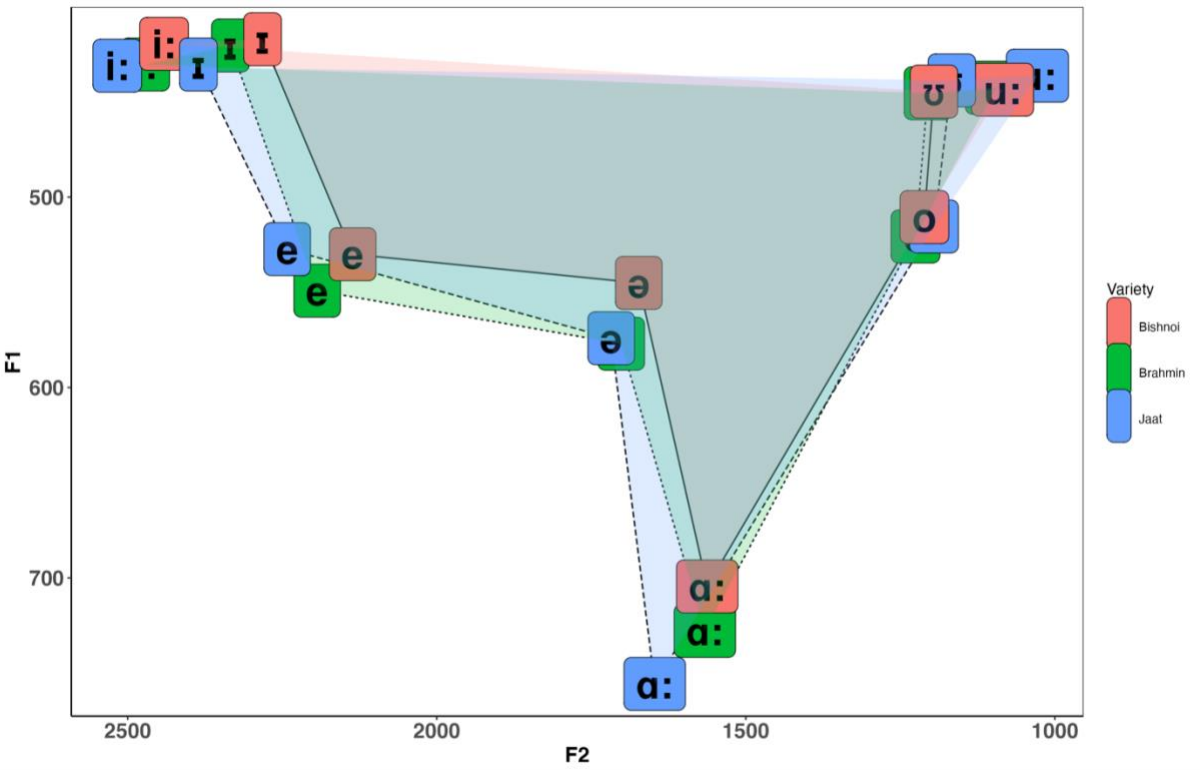


Figure 5.5 A comparative vowel space chart for all three varieties (averaged across all speakers and data types)

The vowel space chart provided a visual representation of the distinctive formant patterns across the Brahmin, Jaat and Bishnoi language varieties. Several key differences stood out:

For the low vowel /ɑ:/, Jaats produce a more fronted and open variant compared to Brahmins and Bishnois, as evidenced by the lower F1 and higher F2 frequencies. This indicates Jaat speakers articulate /ɑ:/with a lowered tongue body position. A similar tendency is seen for the close front vowel /i:/, with Jaats again showing a more fronted and slightly more open vowel quality.

These articulatory patterns are further reflected to a lesser degree in the Jaats' production of /e/, which has a subtly more fronted variant than the other two groups. However, the small standard deviation differences of only 2 Hz and 4 Hz between Jaats and Brahmins and Bishnois respectively for /e/ suggest this is a very modest distinction.

In contrast, Bishnoi speakers demonstrate a consistent backed articulation of front vowels /i:/, /ɪ/ and /e/ compared to Brahmins and Jaats, clearly visible in the more peripheral F2

frequencies. However, this backing pattern does not hold for the back vowels where Bishnois produce variants similar to the other groups. Brahmins generally exhibit formant patterns intermediate between the Jaat and Bishnoi extremes for most vowels.

For the back vowel /o/, all three language varieties share very similar F1 and F2 values indicating a comparable tongue height and backness in articulation. The same holds for the high back vowels /ʊ/ and /u:/, though Jaats show a somewhat more peripheral variant, the standard deviation differences of <10 Hz between groups are negligible.

Mid vowel /ə/ does display a distinct trend, with Bishnois using a more backed and closer variant versus Jaats and Brahmins who occupy similar vowel spaces. This further highlights the Bishnois' tendency to back front vowels but not back vowels.

In summary, while some vowels like /ɑ:/and /i:/ show clear differences pointed out by divergent formant patterns, other distinctions are more modest though still apparent. Together these results demonstrate the subtle but systematic articulatory variations that help distinguish these closely related language varieties.

5.2.2 Vowel space chart for individual data type

Once variety-specific differences were established, the same data was examined across different modes of elicitation. Vowel charts plotted the average performance of each elicitation type per variety. Across varieties, F1 values remained fairly consistent, with moderate variation between groups. However, F2 values displayed more fluctuation across vowels and elicitation types.

For front and mid vowels, those extracted from wordlist data occupied a more open position, with higher F1 and lower F2 values versus story and conversation data. Back vowels showed a similar trend, though F2 differences between modes were less pronounced. Vowels from story data appeared higher and more fronted than wordlist for front and mid vowels. Conversation data was the most fronted for all front and mid vowels. For back vowels, this trend reversed, with conversation data more backed than story data.

This pattern held for the Jaat and Bishnoi varieties. However, some deviations occurred among Brahmin vowels. For instance, /ɑ:/ tokens from wordlists occupied a more backed position than story data, differing from Jaat and Bishnoi, where /ɑ:/ was most fronted across elicitation types.

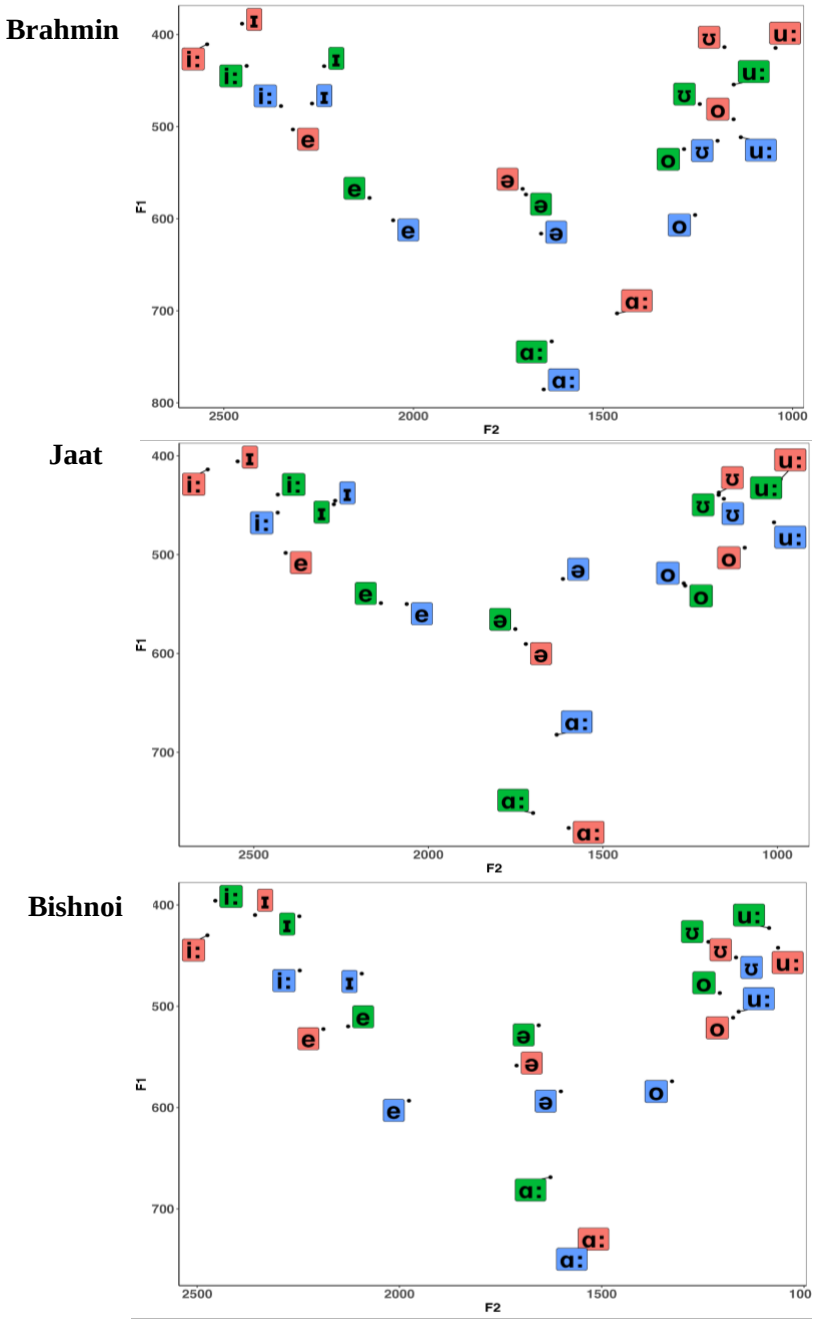


Figure 5.6 Marwari vowel space chart for each mode of data elicitation for every variety (Red = Wordlist, green = story, blue = conversation)

5.3 Within-Formant Features

Chapter 2 discusses the significance of including a within-formant analysis for speaker comparison studies. It highlights the relatively limited research in this area, which consequently means that there are no established set methods to extract these within-formant features automatically. The absence of well-established within-formant analysis studies has presented

numerous challenges in the extraction of these features, particularly in determining the appropriate parameters and settings for their extraction. The next section will elucidate the steps taken to extract within-formant features and will also offer insights into the background reasoning behind the decisions employed during the extraction process.

5.3.1 Extraction of within-formant features

Vowel formant analysis, as demonstrated in section 5.2, revealed inter-variety differences among the Marwari language. To further investigate these differences, a linear mixed-effect regression (lmer) model[7]- based Analysis of Variance (ANOVA)[8] test was conducted on R (Bates et al., 2015; R Core Team, 2023) on initial formant values (James et al., 2021).

This ANOVA analysis aimed to model effects on the formant frequency values using fixed factors of vowels, type, and their interaction, with participants as random factors. The full mixed effects model included the fixed main effects of vowel (8 levels) and type (3 levels: wordlist, story, conversation), their interaction term, as well as by-participant random intercept to account for individual differences among speakers and items.

The test supported the initial vowel space chart-based results. Once the variety-specific differences were confirmed, the subsequent step involved extraction within-formant features (see further p-values table in appendix table 10.2 and 10.3) .

Within the first four formants, eight specific acoustic features were selected for analysis. These features included spectral moments such as the centre of gravity (m_1), standard deviation (m_2),

[7] A Linear Mixed-Effects Regression Model accounts for both fixed and random effects. Fixed effects are predictor variables with a systematic impact on the dependent variable. Random effects introduce variability but are not of primary interest. The dependent variable is the outcome variable to predict or explain. The model is expressed as an equation, similar to a standard linear regression model (James et al., 2021).

[8] ANOVA is a statistical technique used to analyse the variation between groups in a dataset. It divides the total variation into components, assessing the statistical significance of differences between groups. The null hypothesis (H0) states no significant differences between group means, while the alternative hypothesis (Ha) suggests significant differences. ANOVA calculates an F-statistic, a ratio of variance between groups to variance within groups. A large F-statistic indicates significant differences, rejecting the null hypothesis.

skewness (m_3), and kurtosis (m_4), as well as spectral measures including amplitude, relative amplitude, Spectral bandwidth, LPC bandwidth, spectral peaks.

To compute these feature values, a new Praat script was utilised (Harrison, 2021). This script first smoothed the harmonics to enhance the visibility of formants. Afterwards, it automatically identified formants based on previously manually extracted data from the earlier stage. The script's automation relied on finding peaks that were closest to the previously obtained formant data or selecting the values that were as close as possible.

It is important to note that in some instances, the automatically extracted values were not identical to the manually-corrected formant data. This discrepancy arose because, for within-formant analysis, a slightly longer spectra slice was sometimes required. While manually-corrected data could be used on a shorter slice (as they were extracted from two pulses), the automated extraction necessitated a more extended spectral slice to analyse within-formant characteristics. In such cases, the script automatically identified the highest formant peak and marked lower and upper frequencies on both sides for further extraction. These selection points had been predefined in the script, and the script determined the amplitude drop criteria for selection, which could be either +/-3 decibel (dB) or +/-1 decibel (dB). The amplitude drop determined the extraction boundary for the features. The following section will describe each one of these steps in detail by starting with the potential issues that impacted these decisions.

Three main factors had the potential to impact the extraction process. The first factor pertained to the selection of the correct amplitude drop, as it played an important role in determining the boundaries from which the features were to be extracted. Secondly, it was important to make informed choices regarding the appropriate cepstral smoothing values for each formant. Subsequently, the script incorporated formant band values, which aided in creating visual representations of spectral slices. These steps were integral to the accuracy and reliability of the spectral measurement extraction process.

The initial choice was whether to use +/-3dB or +/-1dB as an amplitude drop. The amplitude drop criterion was needed to establish a frequency band width within the formants, i.e., the frequency domain slice of the formant from which within-formant feature values can be extracted. Before proceeding, a brief explanation of why the choice was made and what these values represent is presented.

The 3dB threshold is a widely employed reference point in acoustic studies, particularly for measuring bandwidth. However, it can be adjusted based on the specific research goals. The key concept here is understanding the relationship between changes in sound pressure amplitude and the corresponding impact on intensity/power, which relates logarithmically. Specifically, when the amplitude of a sound wave is decreased by half (or 50%), this constitutes a 3dB drop in the sound intensity and power (Martin, 2021). The decibel scale is logarithmic, meaning a 10dB change corresponds to a 10x increase or 1/10th decrease in power. By extension, a 3dB change up/down signifies a halving/doubling in power.

This demonstrates why the -3dB point is treated as a cut off for meaningful sound power changes - going beyond +/-3dB indicates losing/gaining more than half the power and amplitude, which starts to affect perception of loudness and signal quality. The 3dB level therefore provides a good reference benchmark when evaluating power and intensity shifts in audio signals and waveforms. The second extraction boundary was set for +/- 1dB amplitude drop, which reduces the power by the factor of 20%.

Table 5.4 provides a brief overview of the different amplitude drops and their effects on the sound pressure levels of the signal.

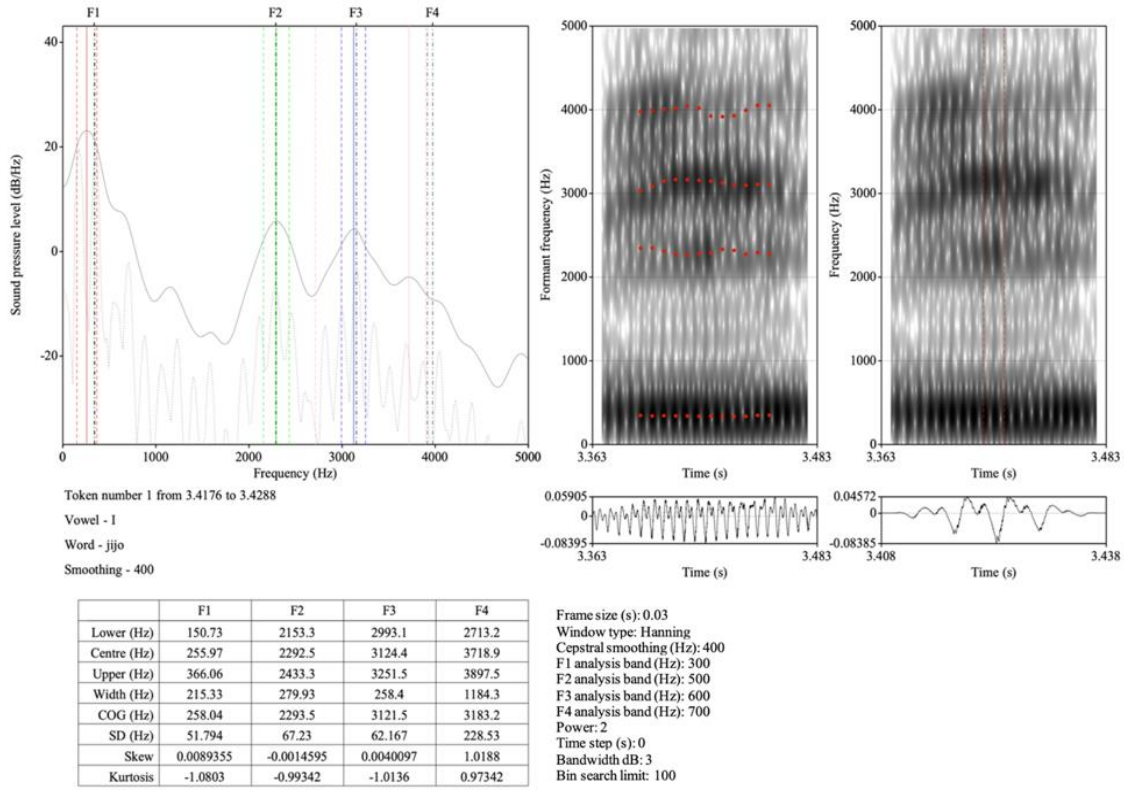
Table 5.4 Power - amplitude relationship at different values

dB	**P_0 * x dB/ 10log (For power/ amplitude)**
0 dB	1 = 100%
+/- 3dB	0.5/2: At -3 dB the power goes by the factor of half and for +3dB it is raised by the factor of 2.
+/- 1dB	0.79 = 21%: At – 1dB the drop in the power is up to 20 %.

Two amplitude drop criteria of +/- 1dB and +/- 3dB from the peak were evaluated during spectral slice extraction to test the potential impact of slice bandwidth on resulting acoustic measurements. The selected drop point determines the upper and lower frequency boundaries for extracting spectral slices and formant contours. Assessing multiple criteria enables evaluation of whether a narrower 1dB or wider 3dB spectral range influences the accuracy and variability of extracted spectral measurements, particularly concentric values within individual formant contours which require precise tracking.

Therefore, both +/- 1dB and 3dB relative amplitude thresholds were examined here for their effect on the precision of extracted within-formant measurements across a large dataset. The

goal was to determine which criteria produced the most reliable spectral slice and fewest outliers or erroneous values that could negatively impact statistical modelling.



	F1	F2	F3	F4
Lower (Hz)	150.73	2153.3	2993.1	2713.2
Centre (Hz)	255.97	2292.5	3124.4	3718.9
Upper (Hz)	366.06	2433.3	3251.5	3897.5
Width (Hz)	215.33	279.93	258.4	1184.3
COG (Hz)	258.04	2293.5	3121.5	3183.2
SD (Hz)	51.794	67.23	62.167	228.53
Skew	0.0089355	-0.0014595	0.0040097	1.0188
Kurtosis	-1.0803	-0.99342	-1.0136	0.97342

Figure 5.7 :An LPC analysis with overlaid spectral slice of a speech segment using pre-determined Praat settings. Formants 1 - 4 estimated by LPC analysis are indicated by the coloured lines traversing the spectral slice. The black dashed lines mark the centre frequencies of formants 1-4 determined from manual analysis for comparison. The red, green, blue and pink dashed lines visually demonstrate the +/- 3dB amplitude drop boundaries automatically extracted around each LPC formant peak to determine the spectral frequency range for formant bandwidth measurements.

The figure displays the LPC (Linear Predictive Coding) spectrum for the first four formants measured from a speech token containing the vowel /ɪ/. Formant estimates from the automated LPC analysis are indicated by the coloured dashed lines traversing the central spectral slice. For comparison, grey dashed lines denote the manually corrected formant frequencies for the same token. The LPC formants align closely to the peaks in the spectral profile, validating the

accuracy of automated measurement. Additionally, dashed boundary lines in red, green, blue and pink illustrate the +/- 3dB amplitude drop thresholds automatically extracted around each LPC formant to determine the upper and lower frequency limits for calculating spectral properties. This demonstrates how the script uses the estimated formant frequencies and then defines a spectral range of analysis based on a standard relative amplitude criterion. Overall, the multiple formant indicators and amplitude thresholds visually convey how the automated script measures resonant frequencies and leverages those estimates to select optimal bandwidth parameters for quantifying spectral shape attributes within each formant energy.

However, there were some notable issues with the extraction process, as evident in Figure 5.7. In this specific extraction, as mentioned earlier, a +/-3dB criteria was employed. Unfortunately, this choice led to some measurement errors, as seen in Figure 5.7, where the lower frequency measurement for F4 falls between F3 and F2 (denoted by the pink line around 2500 Hz). Consequently, when the upper and lower frequencies, defined by the +/-3dB threshold, deviate more than expected, the calculated spectral measures become inaccurate.

The second type of error was that the two adjacent formant peaks sometimes appeared as a single spectral peak because of the small distance between these two adjacent formants for vowels /i:/ and /ɪ/ (F2 and F3) or /u/ and /u:/ (F1 and F2). This was impacted by the smoothing settings and pre-defined formant frequency bands. In these cases, the peak closest to the LPC-estimated formant was taken. This led to the same upper and lower frequencies and bandwidth measures for both formants.

Thirdly, sometimes no upper or lower frequencies could be found in the LPC spectrum because the spectral drop was not steep enough. In these cases, the frequencies were marked as invisible with a symbol of -1.IND. For the purpose of the present study, all tokens with these three kinds of error were eliminated.

Various combinations of amplitude drop (+/- 3dB as 50 % and +/- 1 dB as 20 %) and cepstral smoothing settings were tested to optimise formant extractions. Eight settings combining amplitude drop and smoothing were evaluated. Modifying these settings aimed to avoid the errors generated by extracted spectral feature values falling at unexpected positions.

Any amplitude or smoothing changes caused these issues for various tokens. The goal was to find a set of optimal parameters to extract all features reliably. To achieve this, within-formant feature extraction was conducted using all eight settings presented in Table 5.5.

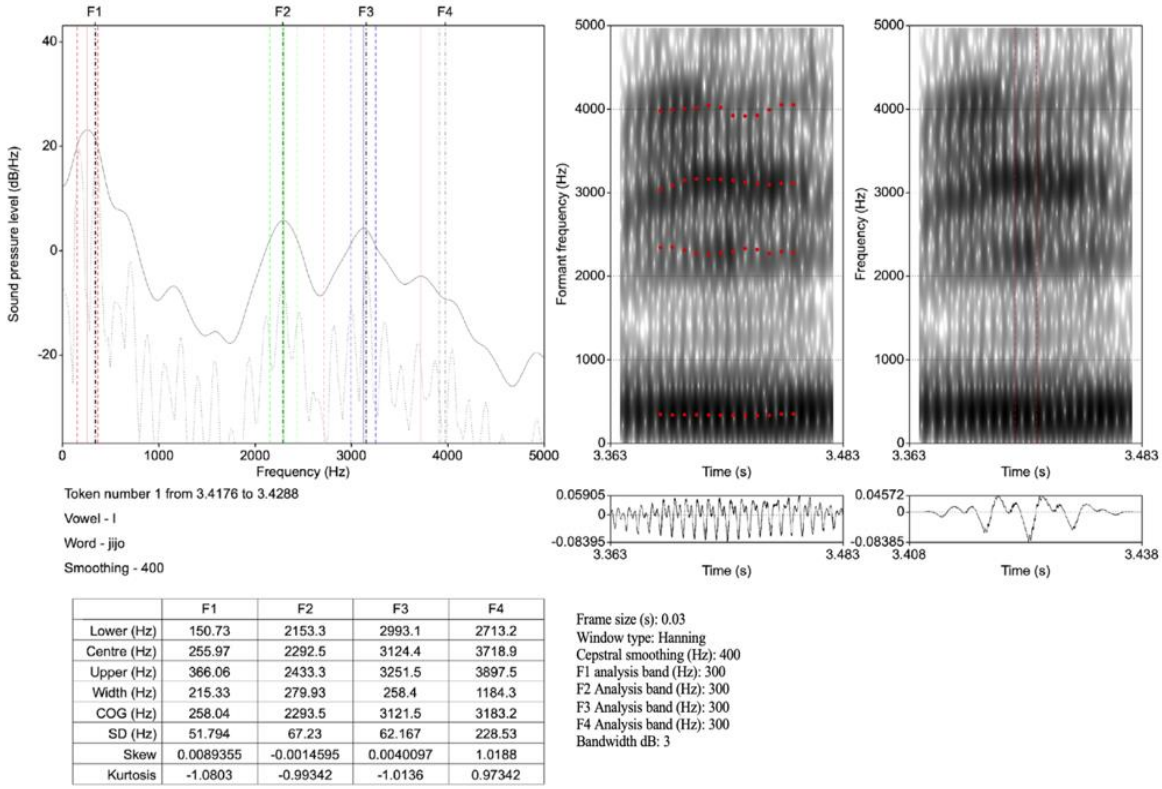
The settings explore modifications to cepstral smoothing and formant bandwidths parameters to balance accurate lower formant values with more flexible higher formant tracking - a combination of techniques suggested to improve LPC precision based on standards and prior optimization research (Simpson, 2009). Settings 1, 2, 7 and 8 reflect common defaults for smoothing and keeping all bandwidths at 300Hz as a consistent baseline. Settings 3-6 loosen the bandwidth ceiling for higher formants F3 (500Hz) and F4 (600Hz) only, while retaining tighter F1 and F2 limits per recommendations. Relaxing upper formant tracking allows better fitting to spectral peaks but risks losing detail. Exploring this trade-off leverages literature guidelines: typical smoothing filters, tightly-constrained F1/F2 bandwidths, and moderately widened F3/F4 settings for performance gains. Collectively these settings test tailored adjustments to maximize LPC accuracy in lower and higher formants using evidence-based techniques.

Table 5.5 Eight settings used for spectral measurement extractions.

Setting	Cepstral smoothing (Hz)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	Bandwidth (+/-Amplitude drop)
1	400	300	300	300	300	3
2	400	300	300	300	300	3
3	400	300	500	600	700	1
4	400	300	500	600	700	1
5	300	300	500	600	700	3
6	300	300	500	600	700	1
7	300	300	300	300	300	1
8	300	300	300	300	300	3

The number of errors was evaluated for each setting with the help of a visual examination. These errors were calculated visually by going through each image extracted from the Praat script. An ANOVA test was also conducted to examine the impact of the settings on the extracted formant features, which showed that setting did have a significant impact on the spectral measurement extracted values. The tested presented p-values <0.05 for each measure tested. The following section provides an overview of the results for each setting and the errors encountered for vowel /ɪ/ as an example.

5.4 Setting 1



	F1	F2	F3	F4
Lower (Hz)	150.73	2153.3	2993.1	2713.2
Centre (Hz)	255.97	2292.5	3124.4	3718.9
Upper (Hz)	366.06	2433.3	3251.5	3897.5
Width (Hz)	215.33	279.93	258.4	1184.3
COG (Hz)	258.04	2293.5	3121.5	3183.2
SD (Hz)	51.794	67.23	62.167	228.53
Skew	0.0089355	-0.0014595	0.0040097	1.0188
Kurtosis	-1.0803	-0.99342	-1.0136	0.97342

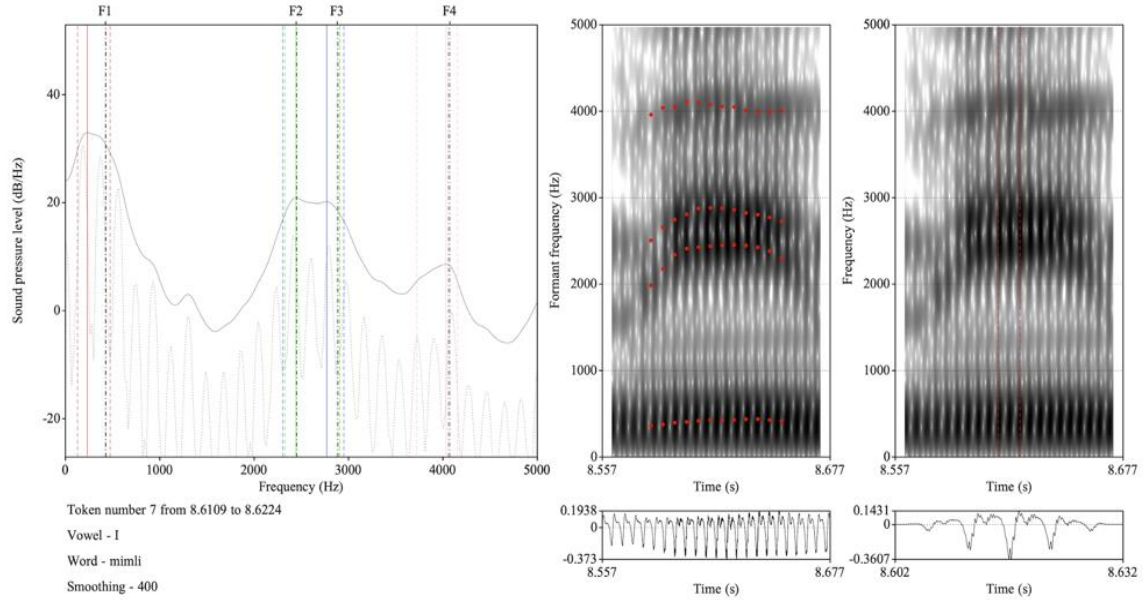
Figure 5.8 LPC and spectrum slice of vowel /ɪ/ for setting 1.

Figure 5.8 is a spectrogram of /ɪ/ extracted from a speaker using setting 1 parameters. This setting produced an error related to inconsistent amplitude drops between formants.

Specifically, for F4, the lower frequency identified using the +/- 3dB amplitude drop falls far to the left of the actual spectral peak, nearly overlapping with the F3 range. This demonstrated how the wider +/- 3 dB slice caused the lower F4 frequency boundary to be improperly identified compared to the actual formant peak location.

This highlights issues that can occur when features are extracted from adjacent formants using different amplitude drop values, leading to inaccurate frequency boundaries and bandwidths.

5.5 Setting 2



	F1	F2	F3	F4
Lower (Hz)	129.2	2325.6	2304.1	3725.2
Centre (Hz)	232.62	2443.2	2768.8	4033.3
Upper (Hz)	473.73	2907	2950	4155.9
Width (Hz)	344.53	581.4	646	430.66
COG (Hz)	293	2606.6	2613.3	3954
SD (Hz)	86.151	156.83	169.44	110.05
Skew	0.12337	0.087196	0.10154	-0.19018
Kurtosis	-1.0649	-1.2024	-1.1552	-1.0413

Figure 5.9 LPC and spectrum slice of vowel /ɪ/ for setting 2.

Setting 2 maintained the +/- 3dB amplitude drop and 400 Hz cepstral smoothing, but the frequency bands were altered from 300 Hz for each formant to 300, 500, 600 and 700 for F1, F2, F3 and F4 respectively. This setting reproduced the same issue as setting 1, suggesting that 400 Hz smoothing with +/- 3dB amplitude drops is not optimal for accurate feature extraction. Keeping the amplitude drop at +/- 3dB amplitude drop while just altering the extraction bands did not resolve the issues with improper formant boundaries.

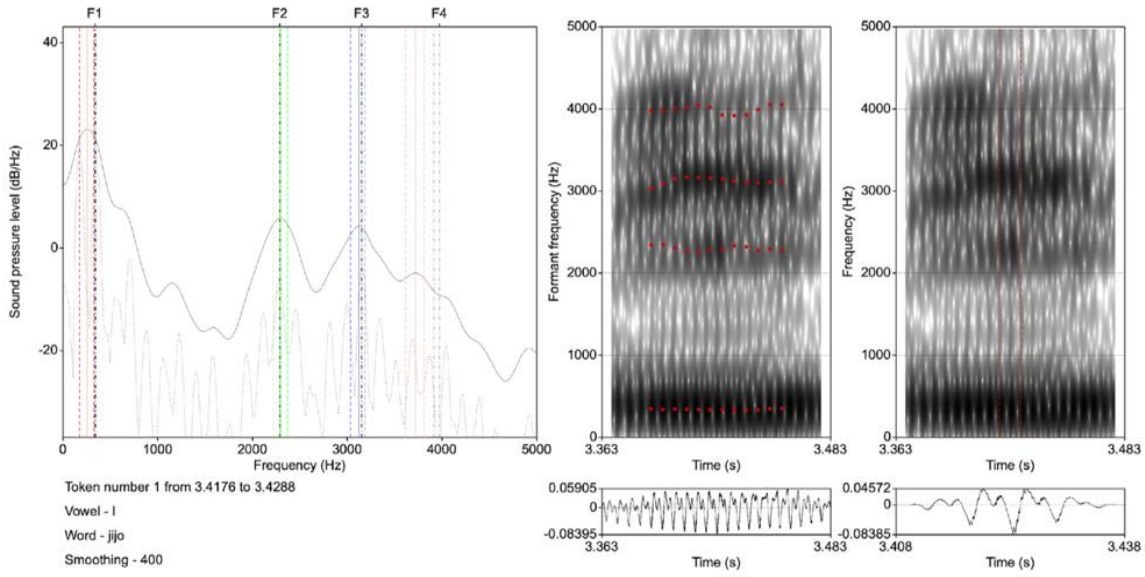
5.6 Setting 3

Figure 5.10 depicts a vowel extracted with setting 3, which reduced the amplitude drop to +/- 1 dB rather than +/-3 dB. The formant analysis bands were also modified back to 300 Hz for each formant.

Setting 3 improved on previous settings by accurately aligning the formant peaks inside the separate analysis bands. The reduced +/-1 dB amplitude drop, on the other hand, generated a

second form of error in which the frequency area between the lower and higher bands was too narrow.

The smoothing settings have an effect on the peak sharpness as well. While setting 3 provided a clean-looking LPC spectrum, the reduced range between boundaries may not adequately capture formant skewness, which was an essential feature to investigate. Statistical testing will be required to see whether the findings for this setting differ too much for skewness and other features; nevertheless, if the results are acceptable, the +/-1 dB setting might markedly minimise extraction inaccuracies.

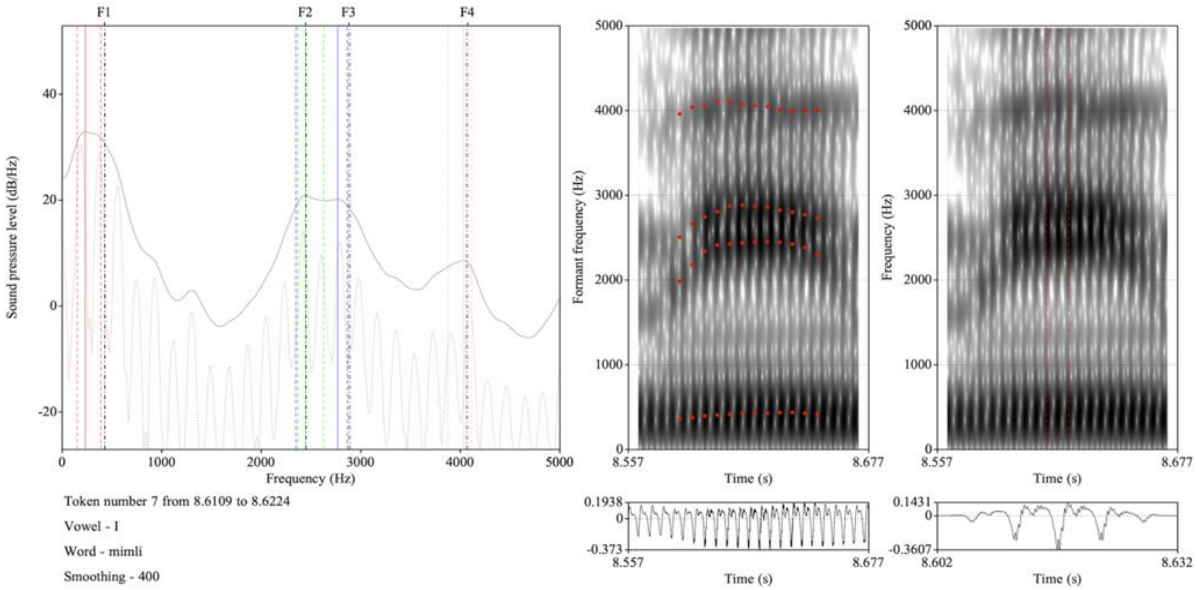


	F1	F2	F3	F4
Lower (Hz)	172.27	2217.9	3036.2	3617.6
Centre (Hz)	255.97	2292.5	3124.4	3718.9
Upper (Hz)	323	2368.7	3186.9	3811.4
Width (Hz)	150.73	150.73	150.73	193.8
COG (Hz)	248.94	2293.3	3113.1	3715
SD (Hz)	35.769	35.846	35.77	48.029
Skew	-0.041971	0.0016717	-0.056837	-0.013689
Kurtosis	-1.2042	-1.2097	-1.2018	-1.1776

Frame size (s): 0.03
Window type: Hanning
Cepstral smoothing (Hz): 400
F1 analysis band (Hz): 300
F2 Analysis band (Hz): 300
F3 Analysis band (Hz): 300
F4 Analysis band (Hz): 300
Bandwidth dB: 1

Figure 5.10 LPC and spectrum slice of vowel /ɪ/ for setting 3.

5.7 Setting 4



	F1	F2	F3	F4
Lower (Hz)	150.73	2368.7	2347.1	3876
Centre (Hz)	232.62	2443.2	2768.8	4033.3
Upper (Hz)	387.6	2627.1	2863.9	4112.8
Width (Hz)	236.87	258.4	516.8	236.87
COG (Hz)	268.52	2494	2598.2	3996.8
SD (Hz)	60.254	66.963	142.31	60.383
Skew	0.036395	0.097087	0.08272	-0.061045
Kurtosis	-1.1768	-1.1808	-1.2331	-1.1733

Frame size (s): 0.03
Window type: Hanning
Cepstral smoothing (Hz): 400
F1 analysis band (Hz): 300
F2 analysis band (Hz): 500
F3 analysis band (Hz): 600
F4 analysis band (Hz): 700
Bandwidth dB: 1

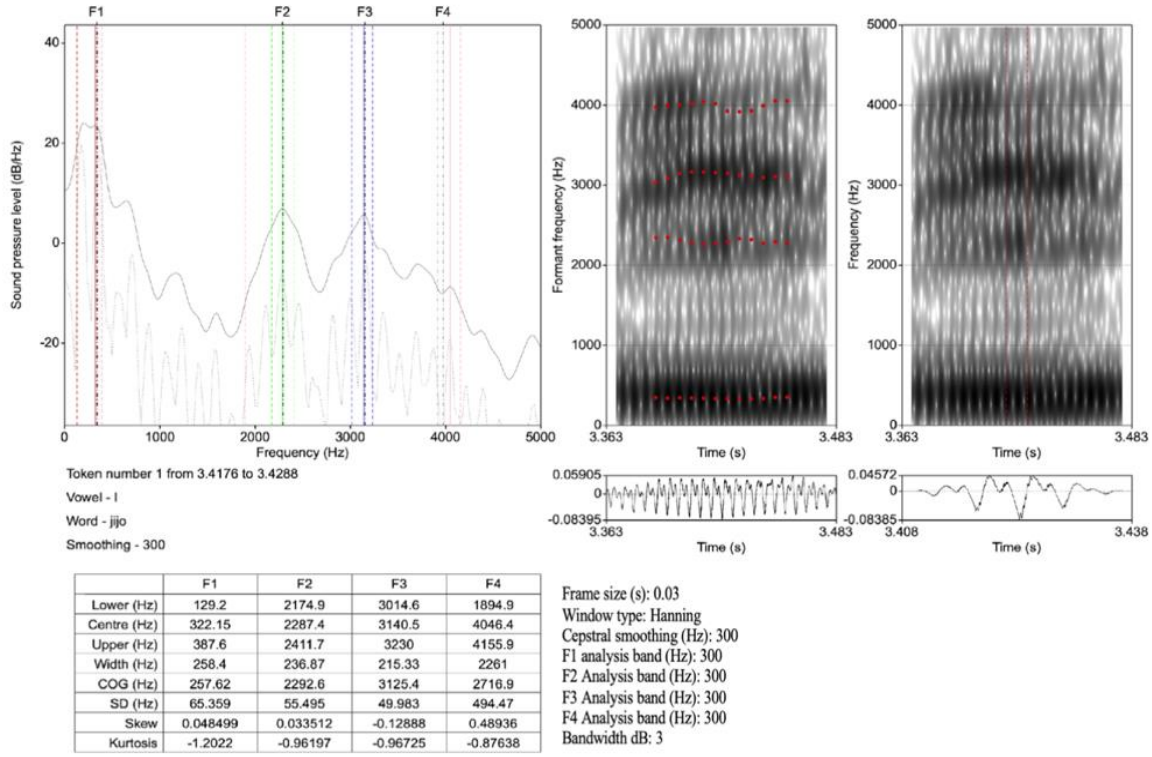
Figure 5.11 LPC and spectrum slice of vowel /ɪ/ for setting 4.

Setting 4 used 400 Hz cepstral smoothing, which greatly impacted the sharpness of the formant peaks. While individual peaks are still visible in Figure 5.11, they appear much flatter compared to previous settings.

The dashed lines marking the frequency boundaries are approximately aligned with the peaks. However, the flatter formant peaks make it harder to identify the true highest point of each peak. This is evident in the F2 and F3 regions of the Figure 5.11.

The 400 Hz smoothing combined with the predefined formant bands created this spectrum with poorly defined, flat peaks. While the frequency boundaries are placed reasonably on the peaks, the reduced sharpness can affect precise spectral moment extractions.

5.8 Setting 5



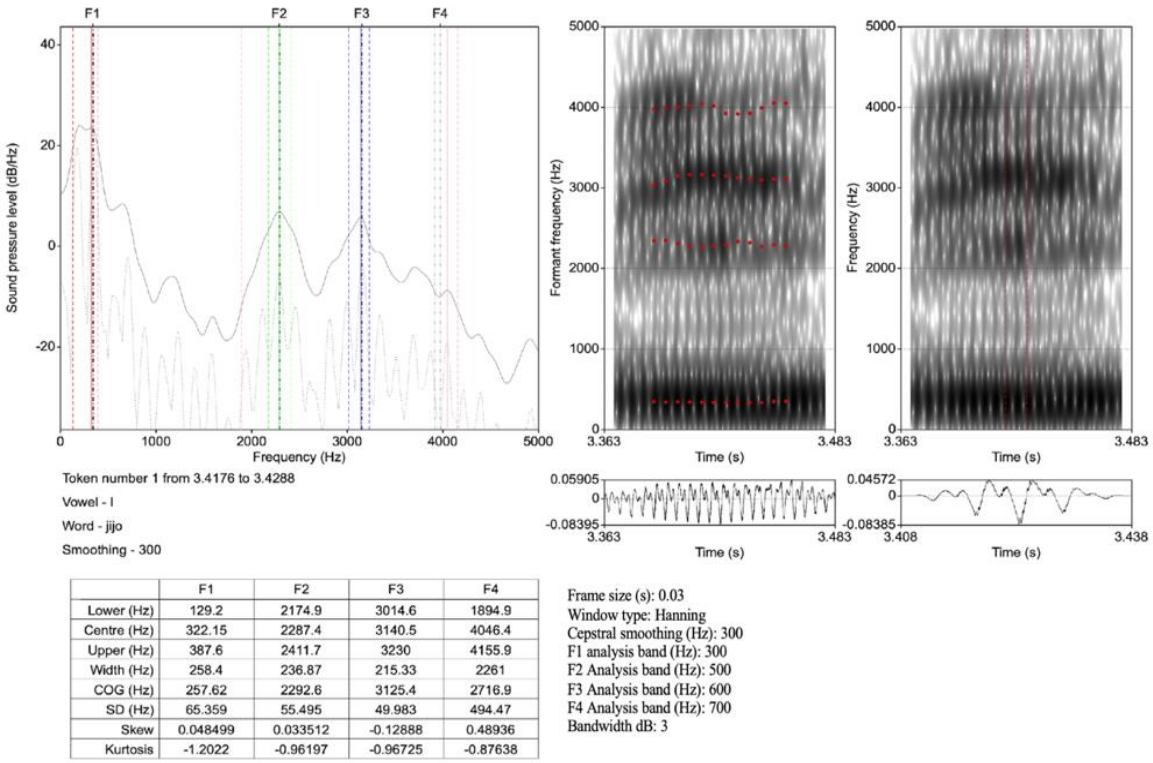
	F1	F2	F3	F4
Lower (Hz)	129.2	2174.9	3014.6	1894.9
Centre (Hz)	322.15	2287.4	3140.5	4046.4
Upper (Hz)	387.6	2411.7	3230	4155.9
Width (Hz)	258.4	236.87	215.33	2261
COG (Hz)	257.62	2292.6	3125.4	2716.9
SD (Hz)	65.359	55.495	49.983	494.47
Skew	0.048499	0.033512	-0.12888	0.48936
Kurtosis	-1.2022	-0.96197	-0.96725	-0.87638

Figure 5.12 LPC and spectrum slice of vowel /ɪ/ for setting 5.

Cepstral smoothing was lowered to 300 Hz for setting 5, resulting in errors where the extraction limits drifted further away from the real formant energy area. F4, for example, dropped below F2 in the lower frequency range. As with setting 1, this option produced a larger number of errors, indicating that settings with +/-3dB produce a higher number of first kinds of errors.

5.9 Setting 6

For setting 6, the individual frequency bands were changed from 300 Hz for each formant to 300 Hz, 500 Hz, 600 Hz and 700 Hz for F1 to F4 respectively. Same as setting 5, this setting shows a first kind of error. As shown in Figure 5.13, the F4 spectral measurements would be affected because of where they are extracted from, i.e., the +/-3 dB drop on the lower frequency is at the wrong place for F4.



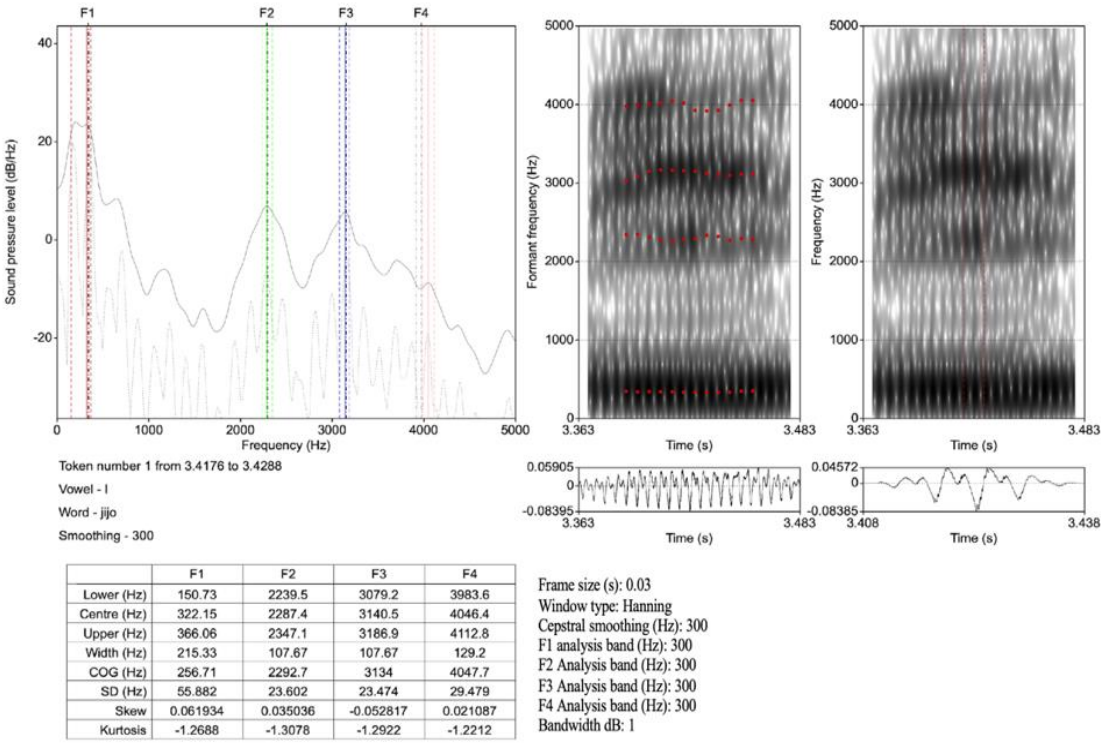
	F1	F2	F3	F4
Lower (Hz)	129.2	2174.9	3014.6	1894.9
Centre (Hz)	322.15	2287.4	3140.5	4046.4
Upper (Hz)	387.6	2411.7	3230	4155.9
Width (Hz)	258.4	236.87	215.33	2261
COG (Hz)	257.62	2292.6	3125.4	2716.9
SD (Hz)	65.359	55.495	49.983	494.47
Skew	0.048499	0.033512	-0.12888	0.48936
Kurtosis	-1.2022	-0.96197	-0.96725	-0.87638

Frame size (s): 0.03
Window type: Hanning
Cepstral smoothing (Hz): 300
F1 analysis band (Hz): 300
F2 Analysis band (Hz): 500
F3 Analysis band (Hz): 600
F4 Analysis band (Hz): 700
Bandwidth dB: 3

Figure 5.13 LPC and spectrum slice of vowel /ɪ/ for setting 6.

5.10 Setting 7

Figure 5.14 displays the results for setting 7, which used same smoothing and frequency band parameters as setting 6, but the amplitude drop was reduced to +/-1 dB. Similar to settings 3 and 4, this created a narrower slice of the spectrum compared to +/- 3dB drop. In the figure, there are two visible closely spaced F1 peaks. The proximity of these adjacent F1 formants peaks could potentially impact the accuracy of feature extraction using the narrower +/- 1 dB (as presented for setting 3 and 4, this could impact the spectral moment values, especially the skewness).



	F1	F2	F3	F4
Lower (Hz)	150.73	2239.5	3079.2	3983.6
Centre (Hz)	322.15	2287.4	3140.5	4046.4
Upper (Hz)	366.06	2347.1	3186.9	4112.8
Width (Hz)	215.33	107.67	107.67	129.2
COG (Hz)	256.71	2292.7	3134	4047.7
SD (Hz)	55.882	23.602	23.474	29.479
Skew	0.061934	0.035036	-0.052817	0.021087
Kurtosis	-1.2688	-1.3078	-1.2922	-1.2212

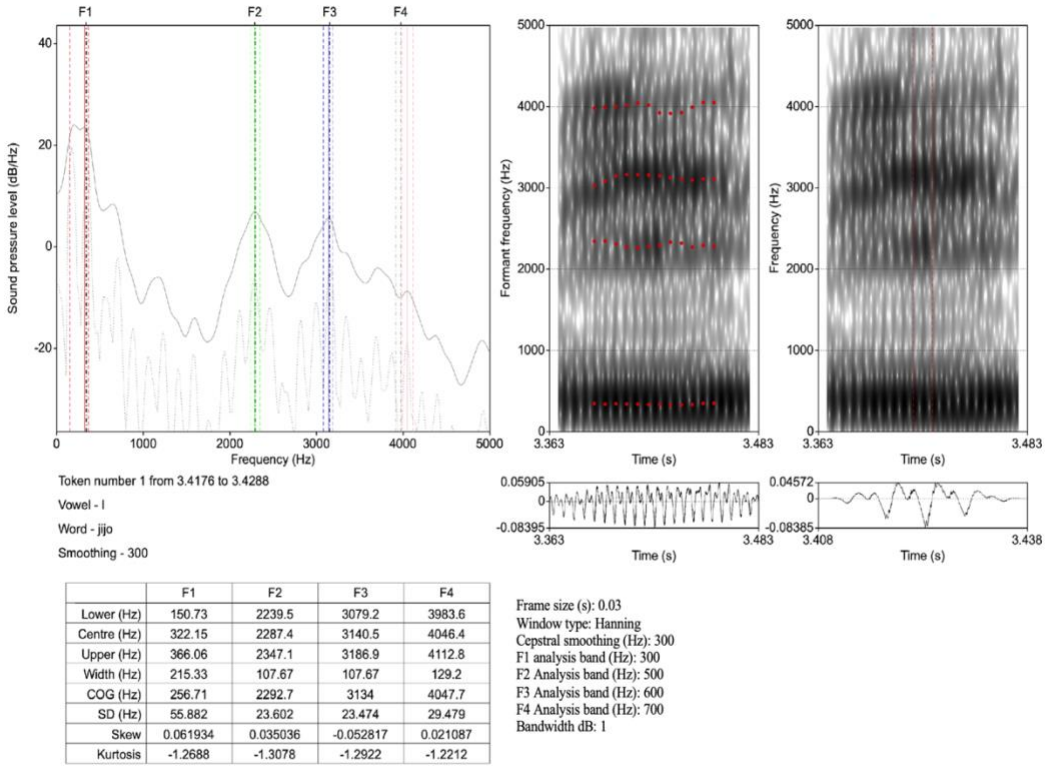
Figure 5.14 LPC and spectrum slice of vowel /ɪ/ for setting 7.

5.11 Setting 8

Setting 8 differed from setting 7, as the frequency bands were adjusted to 300, 500, 600 and 700Hz for F1-F4 (Figure 5.15). This aimed to avoid errors from setting 7 which had the same amplitude drops. As seen in Figure 5.15, the formant boundaries are still very narrow. This setting also provided two F1 peaks in close proximity to each other.

The errors that occurred with visual analysis were tabulated and are presented in Table 5.6. This table quantified the various extracted errors encountered with each setting combination based on inspecting the plots.

Based on the error analysis, setting 7 and 8 with +/1 dB amplitude drops had the fewest errors. For +/- 3 dB drop setting 6 had the lowest errors.



	F1	F2	F3	F4
Lower (Hz)	150.73	2239.5	3079.2	3983.6
Centre (Hz)	322.15	2287.4	3140.5	4046.4
Upper (Hz)	366.06	2347.1	3186.9	4112.8
Width (Hz)	215.33	107.67	107.67	129.2
COG (Hz)	256.71	2292.7	3134	4047.7
SD (Hz)	55.882	23.602	23.474	29.479
Skew	0.061934	0.035036	-0.052817	0.021087
Kurtosis	-1.2688	-1.3078	-1.2922	-1.2212

Figure 5.15 LPC and spectrum slice of vowel /ɪ/ for setting 8.

Table 5.6 Percentage of errors determined by visual inspection of the imaged of each LPC spectral slice extracted for the measurements for individual settings.

Setting	Smoothing	F1	F2	F3	F4	Amplitude	Error Rate
1	400	300	300	300	300	3	38.9
2	400	300	500	600	700	3	38.2
3	400	300	300	300	300	1	25.56
4	400	300	500	600	700	1	19.11
5	300	300	300	300	300	3	21.14
6	300	300	500	600	700	3	23.06
7	300	300	300	300	300	1	6.9
8	300	300	500	600	700	1	6.9

In conclusion, these settings presented following results:

1. Both types of amplitude drop have their own set of limitations. The extraction boundaries are often shifted further away from the real formant frequency ranges for +/- 3dB, especially for higher formants. For +/- 1 dB, however, a significantly narrower spectral slice is produced to extract within-formant features. The narrowness of the slice

may make obtaining the spectral moment challenging (this hypothesis will be further tested with the help of linear mixed-effect regression model-based analysis).

2. Based on a visual evaluation of each token, the frequency band modifications had no significant influence on the extraction process.

3. Cepstral smoothing affected the sharpness of formant peaks. Which could further impact the spectral feature values.

Based on these findings, settings 6 and 8 were chosen for future investigation since they had the fewest errors in each amplitude drop category. As these smoothing settings had no discernible effect on the visual examination, the settings with default parameters were selected, i.e., 300 (F1), 500 (F2), 600 (F3), and 700. (F4).

The relevance of the two different settings on extracted features was assessed with the help of lmer model-based ANOVA tests in RStudio (R Core Team, 2023). These tests examined two different models, one with settings as a dependent variable (a factor) and one without it.

Preliminary visual examination revealed that settings might have a direct influence on spectral moment values. Consequently, the extracted values of spectral moments were first tested to assess if settings significantly impacted them (If no influence on spectral moments is discovered, the test will be performed on spectral measurements.). Table 5.7 presents the p-values[9] obtained from the ANOVA testing of one of these two models with +/- 3dB amplitude drop. Similar results were obtained for +/- 1 dB amplitude drop as well, for which p-values were <0.5. The table makes it clear that settings indeed played a substantial role in the extraction of spectral moments, as evidenced by the highly significant p-values (<.05) (since the test already demonstrated significant impact on spectral moments, additional testing on spectral measures was deemed redundant, as during the analysis both types of features are needed).

[9] The p-value in a linear mixed-effects regression model (LMER) indicates the significance of fixed effects in the model. A small p-value indicates strong evidence against the null hypothesis, suggesting at least one fixed effect has a significant impact on the outcome variable. A large p-value suggests insufficient evidence to reject the null hypothesis. Researchers use p-values to determine the significance of predictors, but consider context, research question, and potential issues with multiple testing (James et al., 2021).

Setting 6 (+/-3dB amplitude drop) was selected for further analysis despite encountering more errors than setting 8. A +/- 3dB drop was judged to capture spectral moments more accurately by providing a larger spectral slice and higher energy distribution. As there is no literature available for such analysis, a judgement call was made to use the traditional +/- 3dB amplitude drop, which provided the least number of errors. Additionally, setting 6 was deemed to reduce the issue of limited extraction boundaries that occurred with Setting 8. However, this was a judgment call, and in principle, the analysis could be repeated and tested with Setting 8 as well.

Table 5.7 Error rates identified for individual settings.

Feature	Wordlist	Story	Conversation
F1- m_1	<< .0001	0.0009	0.0813
F2- m_1	0.9215	0.0170	0.8616
F3- m_1	0.0085	0.8249	0.3094
F4- m_1	0.0036	0.0397	0.1382
F1- m_2	<< .0001	<< .0001	<< .0001
F2- m_2	<< .0001	<< .0001	<< .0001
F3- m_2	<< .0001	<< .0001	<< .0001
F4- m_2	<< .0001	<< .0001	<< .0001
F1- m_3	<< .0001	<< .0001	1<< .0001
F2- m_3	0.2558	0.0131	0.4912
F3- m_3	0.1833	0.8688	0.4755
F4- m_3	0.6446	0.1904	0.8617
F1- m_4	<< .0001	<< .0001	<< .0001
F2-m_4	<< .0001	<< .0001	<< .0001
F3-m_4	<< .0001	<< .0001	<< .0001
F4-m_4	<< .0001	<< .0001	<< .0001

The next stage of analysis involved conducting a statistical analysis of these within-formant measurements to determine if vowels, varieties and modes of data elicitation affect the feature extraction process.

5.12 Assessing the Impact of Varieties and Vowels

Once the optimal settings were determined, a series of lmer model-based ANOVAs were conducted for each feature individually. The aim of these ANOVAs was to assess if variety or vowel had a significant impact on the measurements. Lmer models were created with variety and vowel (*Var+Vow*) as dependent variables (factors) and participant as random variable. To evaluate significance, the *Var+Vow* model was compared to three separate models: one with variety alone (*Var*) as a dependent variable, one with vowel alone (*Vow*) as a dependent

variable, and one with the interaction between variety and vowel (*Var*Vow*) as dependent variables. ANOVA tests were used to determine the significance level for each model.

The models were tested for four spectral moments extracted from formant mid-points for F1-F4: centre of gravity (m_1), standard deviation (m_2), skewness (m_3) and kurtosis (m_4). Additionally, the models were tested for selected spectral measures: amplitude (A), spectral peak (SP), LPC bandwidth (LB), spectral bandwidth (SB) and the difference between amplitude values of each formant.

For spectral moments, the p-values for m_1 for each formant in all models were found to be < 0.05, indicating significance. There were two exceptions: F4 *Var*Vow* and F3 *Var*Vow* in the conversation data, which had p-values > 0.05, suggesting that, for these two models, the interaction between variety and vowel did not significantly affect the values of m_1.

The p-values were generally higher for m_2 compared to m_1, resulting in a lower significance level. For the higher formants (F3-m_2 and F4-m_2), *Var*Vow* had a p-value >0.05, for all three data elicitation types, implying that the interaction of vowels and varieties did not have a substantial impact on the values of m_2. The was also the case for F2 of story and conversation data.

The significance of variety reduced noticeably for the next two spectral moments, with only F3m_3 and F1m_4 for wordlist data having significant values. For m_3, none of the models showed significant p-values except for wordlist *Vow* (F1, F2, F4) and *Var*Vow* (F1) in wordlist data, suggesting that variety and vowel do not significantly influence skewness.

For m_4, significant values were observed for *Vow* in wordlist data, but only one significant for story (F3-m_4) and conversation (F2-m_4) data. The *Var model* was only significant for F1-m_4 in wordlist and story data, suggesting that variety does not substantially impact kurtosis.

In summary, both variety and vowel were highly significant for the first two spectral moments. However, the interaction between them, while still significant, had a lesser effect compared to treating them as individual dependent variables. The p-values for each spectral moment extracted from individual formants are presented in Appendix table 10.2.

The results of lmer model-based ANOVA conducted on the spectral measures indicated that, for wordlist data, all three models performed significantly well (the p-values < 0.05), with only a few exceptions. These exceptions were observed in the variety-dependent models of spectral

measures (specifically, LB1, SB2, SB3, and A4), suggesting that, in the context of wordlist data, the null hypothesis (that variety did not have any impact on spectral measure values) was rejected. Interestingly, some spectral measures (e.g., A1, A2, A4, A4-A2, LB1, LB4, and SP4) showed less significant interactions between variety and vowels, but these measures were not significantly impacted by variety alone either. This implied that the linguistic variety itself does not have a substantial influence on the values of selected spectral measures.

However, the mode of data elicitation had a notable impact on the spectral measures. The significance of p-values decreased for story and conversation data, particularly affecting the values of spectral measures for the conversation data (see Table 10.3 in the Appendix for further information). While the number of insignificant values increased for variety and the interaction of variety and vowel for story data, this number significantly escalated for both models in the context of conversation data.

In summary, vowels remained a significant factor in determining the values of spectral measures across all three modes of data elicitation.

5.12.1.1 Classification method

A linear discriminant analysis (LDA) is a statistical technique that classifies data into two or more groups (Fisher, 1938; Martinez & Kak, 2001). Unlike other dimensional reduction techniques such as principal component analysis (PCA) which are unsupervised, LDA is a supervised technique that searches for the underlying vectors or a linear combination of features that best separates the groups. It also differs from PCA as it identifies features that maximise the separation between classes and orders them by their discriminative powers, whereas PCA finds the set of features that captures the most variation in the data and orders them by their importance. LDA was selected over PCA because the present study requires classification tasks rather than exploratory data analysis. LDA is based on two main assumptions: data is normally distributed, and each class has equal covariance matrices. It is often used in image classification, text classification and biomedical engineering. For speech recognition, LDA can be used to classify speech into different speakers or different emotions (Boedeker & Kearns, 2019; Martinez & Kak, 2001).

5.13 Conclusion

Data processing involved several steps, which can be summarised as:

Step 1 The initial analysis of sound files included isolating specific speech sounds and normalising peak levels to 2.0 dBFS through peak normalisation.

Step 2 Vowel sounds from CVC syllable structure were then extracted from Praat.

Step 3 Formant values were logged using a Praat script and subsequently manually corrected with a visual examination.

Step 4 For each participant, ten tokens of eight different vowels were extracted totalling 80 tokens per participant.

Step 5 A vowel space chart was generated to analyse the average space occupied by each variety.

Step 6 Spectral features were extracted under various settings, including +/-3dB and +/1dB amplitude drops. These settings were further adjusted based on the cepstral smoothing and formant band values.

Step 7 Spectral feature extraction was performed for eight different settings, and two settings with the fewest errors were selected for further testing. These settings included +/-3dB and +/-1dB amplitude drop, each with the same cepstral smoothing of 300 Hz and formant bands of 300 (F1), 400 (F2), 500 (F3) and 600 (F4).)

Step 8 A linear mixed model ANOVA was conducted to assess whether the chosen settings had a significant impact on the extracted feature values. The ANOVA results indicated significant p-values, suggesting that the settings did indeed affect the feature values. The setting with the +/-3dB amplitude drop was selected for further analysis.

Step 9 The final stage of data processing involved examining the influence of vowel, variety and mode of data elicitation on spectral features.

These steps constitute the comprehensive data processing pipeline employed in the study. The subsequent section of the study will be divided into four articles, each focusing on different parameters that were examined.

The first article investigates the within-formant spectral moment analysis, while the second article assesses the effects of spectral measure analysis.

The third article will present findings from a combined analysis incorporating both spectral moment and measures, showcasing models with various combinations of these factors.

6 Chapter 6 Presented as Article 1

6.1 Research Degree Thesis Statement of Authorship

University of York

York Graduate Research School

Candidate name	Nikita Suthar
Department	Language and Linguistic Science
Thesis title	Within-formant spectral feature analysis for forensic speaker discrimination casework: A study of 45 Marwari monolinguals from Bikaner, India

Title of the work (paper/chapter)	Primary spectral moments of the first four vowel formants as a source of speaker discriminant information	
Publication status	**Published**	
	Accepted for publication	
	Submitted for publication	*
	Unpublished and unsubmitted	
Citation details (if applicable)	Suthar, Nikita and French, John Peter, Primary Spectral Moments of the First Four Vowel Formants as a Source of Speaker Discriminant Information. Available at http://dx.doi.org/10.2139/ssrn.4581148	

Description of the candidate's contribution to the work	Conceptualisation, literature review, data collection and analysis, writing and manuscript preparation, citation, and references

Percentage contribution of the candidate to the work	90%
Signature of the candidate	Nikita Suthar
Date (DD/MM/YY)	25th September 2023

Co-author contributions*

By signing this Statement of Authorship, each co-author agrees that:

(i) the candidate has accurately represented their contribution to the work;

(ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co-author	Prof. Peter French
Contact details of co-author	peter.french@york.ac.uk
Description of the co-author's contribution to the work**	Part of conceptualising, contribution to interpretation of findings, editing
Percentage contribution of the co-author to the work	10%
Signature of the co-author	
Date (DD/MM/YY)	27th Sept 2023

6.2 Title: Primary Spectral Moments of the First Four Vowel Formants as a Source of Speaker Discriminant Information

6.3 Abstract

This study assesses the potential benefit of analysing spectral moments (SMs) for forensic speaker comparison work. An analysis of the first four primary SMs (extracted from the first four vowel formants retrieved at midpoint), i.e., centre of gravity (m_1), standard deviation (m_2), skewness (m_3) and kurtosis (m_4), is presented. Data was collected from forty-five female speakers of Marwari (an Indo-Aryan language from northwest India). The analysis was conducted on eight different vowels (ten tokens per vowel) obtained from three different modes of elicitation (wordlist, story, conversation). The first part of the study discusses the significance of spectral moment analysis (SMA) for distinguishing between different varieties of language. This analysis is followed by a linear discriminant analysis conducted on the SM values to identify the best possible spectral moment or combination of moments, for discriminating between different speakers of the same variety. Results showed that the variety, vowel, and mode of data elicitation significantly impacted the SMA. Linear discriminant analysis results also suggested that m_1 and m_2 performed significantly better than the other SMs in classifying speakers, increasing the correct classification rate of participants to 9 times more than the chance level.

***Keywords**: Vowel formants, spectral moment analysis, centre of gravity, standard deviation, skewness, kurtosis.*

6.4 Introduction

The analysis of vowel formants and their importance in forensic speaker comparison work has been the focus of multiple studies (e.g. Burris et al., 2014; Cao & Dellwo, 2019; Fleischer et al., 2015; Gonzalez-Rodriguez, 2011; McDougall, 2006; Nolan & Grigoras, 2005). Certain studies have suggested that the higher formants of vowels carry speaker-specific information (e.g. Cao & Dellwo, 2019; Jessen, 1997). Milenkovic and Forrest (1988) pointed out that, although formant centre frequencies may be a reliable measure for speaker discrimination, it

can be challenging to accurately identify a formant peak manually and spectral moment analysis (SMA) can help overcome this problem.

Spectral moments (SMs) have also been described as an important and reliable tool for identifying the place of articulation for fricative consonants and age-specific differences based on the acoustic cues in higher frequency ranges (e.g. Forrest et al., 1988; Jongman et al., 2000; Maniwa et al., 2009; Nissen & Fox, 2005; Nittrouer, 1995).

In addition to this, SMA has been applied in some vowel perception studies (e.g. Chistovich & Lublinskaya, 1979; Milenkovic & Forrest, 1988) and speaker discrimination studies (e.g.Eriksson, Cepeda, Rodman, McAllister, et al., 2004; Rodman et al., 2002) but for the most part, it remains under-researched.

This paper hypothesises SMs of vowel formants, especially the higher formants, might carry individual speaker information. Grounds for this view are to be found in Rodman et al.'s (2002) contention that SMs represent glottal pulse shape and are, therefore, dependent on the physiology of an individual's phonatory organs and supralaryngeal vocal tract.

The present research, like that of Rodman et al. (2002), mainly uses text-independent speech data and analyses the SMs of vowels' first four formant spectra.

In addition to investigating whether SMA is effective in distinguishing between individual speakers, we also address the questions of whether it performs better for some varieties of a language rather than others, whether its effectiveness is influenced by the vowels from which the SMs are extracted, and whether its performance is affected by mode of speaking/speaking style.

In the sections below, we describe our methods of data collection and how spectral moments were determined and calculated before progressing to the various analyses.

6.5 Data Collection

The Marwari language on which the study is based is an Indo-Aryan language spoken mainly by the members of the Marwari community (also called Marvari, Marvadi and Marwadi) residing in the north-western areas of Rajasthan (a state in northwest India).

Three different caste-based varieties of Marwari were recorded. Caste is an occupation-based classification system in Hinduism for society. The first variety is spoken by members of the

Brahmin caste who are primarily associated with education or theology. Brahmin is the highest Varna of the Hindu caste system (Jain, 1979). (A Varna in Hinduism is defined as a social class within the hierarchical caste system.) The second variety spoken by the Jaat caste is predominantly involved with farming and herding cattle in Rajasthan. It belongs to the 'Vaisya' Varna (caste primarily associated with the business section) or 'Shudra' Varna (caste primarily associated with the working section of the Hindu caste hierarchy). The categorisation of Jaats between Vaisya and Shudra depends on the state or region of the country they belong to. The third variety is Bishnoi. The Bishnoi caste or community is the newest of the three castes. The word community is used here because Bishnoism started to abolish castism in India. However, overtime members of this community started identifying themselves as a member of Bishnoi caste. Bishnoi is a community of nature worshipers who originated in the 15th century with their Guru Jambheshwarji, who presented twenty-nine rules for his followers. These twenty-nine rules, or as they are called /bi:s/ (twenty)+/noi/ (nine), later created the name of the community (Jain, 2010). Many of the present Bishnoi speakers have roots in the Jaat before the 15th century. Their respective cultural and social values are very similar (Jain, 2010).

To exclude regional variation as a variable, the recordings were collected from life-long resident female monolingual speakers from the Bikaner district in Rajasthan.

Forty-five participants were recruited, with fifteen speakers for each variety. All participants were born and raised in the Bikaner district. They are all above forty with a mean age of 50.68 (range 40-84, standard deviation = 8.03). Each participant was asked about their educational qualifications and linguistic competence. The primary basis for selecting these participants was their monolingualism. All participants were naïve to the specific research questions but understood the general purpose of the study. Speakers from the Bishnoi and Jaat varieties predominantly reside in the rural areas of the district. Brahmins, on the other hand, live in urban areas. The researcher visited the participants' houses and made the recordings in the quietest non-echoic room.

Recordings were of spontaneous and non-spontaneous speech. Three modes of data collection were employed: The first mode was a wordlist, i.e., a list of everyday words represented in the Devanagari script, and they were asked to read those words. Some speakers could not read, so to accommodate them, an informant (who was an assistant and did not act as a participant) from their variety was asked to read, and the participants repeated after them. It is noted here that

this method might have influenced participants to converge or diverge their speech based on the informant's speech (Giles, 1973; Pardo et al., 2022). The initial assumption was that there would be a higher degree of convergence happening between the informant and the participant because of two reasons: the interlocutor was from a different age group which could result in difference in language use (Babel, 2009, 2012; Cao, 2018) and both participant and informant were very similar to each other (they both belonged to same family and were of the same gender) (Earnshaw, 2021; Pardo et al., 2018). However, it was noted that every time there was a difference in the pronunciation between informant and the participant, participants always stated that their forms are correct as they are older, thus suggesting more divergence than convergence (only the participant's forms were accepted). This was determined because the researcher was present during the data collecting procedure and inquired why participants said their form was correct or different. The studies mentioned earlier focused on a lab-based semi-spontaneous speech where participants were educated and could read. In a speech community such as Marwari females from semi-urban populations, where participants were uneducated, sometimes compromises must be made. In any case, it doesn't not seem to have had a negative impact on results (see below).

The second mode was a picture description task, i.e., participants were shown a picture of local deities and were asked to narrate a story associated with them. The third method was a normal conversation, where two participants were put together and asked to discuss a topic of their choice, or a topic chosen from a list provided. All three modes of data collection were conducted at one sitting. Participants were recorded under controlled conditions (quiet-non-echoic room) using a high-quality digital recording device, 'Zoom H4n Handy Recorder'[10] (files: .wav format; 44.1 kHz sampling rate; 16-bit depth). This recorder had built-in microphones that could be adjusted to 90 or 120 degrees, as required. The recordings were made on two different channels on the recorder. The recorder was positioned 25 centimetres from each participant's mouth on a tripod. The recorder's microphones were adjusted to 120 degrees for both channels depending on the participant's position: the controlled conditions involved a quiet space in their house, with minimum distraction or noise.

[10] Specifications: https://www.zoom.co.jp/sites/default/files/products/downloads/pdfs/E_H4nSP_0.pdf

Phonetically, the three caste-based varieties exhibit some apparent differences. As presented in, Figure 6.1, the vowel space occupied by the three varieties shows that the Jaat variety has more fronted and open vowels than the other two. Eight different vowels were selected for this study. These vowels were selected because of their presence in each selected variety. It should be noted that these varieties have more vowels, including diphthongs and the sole rationale for selecting these vowels was that they were the only common monophthongal vowels present in all three varieties. The eight vowels selected here are:

[iː], [ɪ], [e], [ə], [aː], [o], [uː], [ʊ]

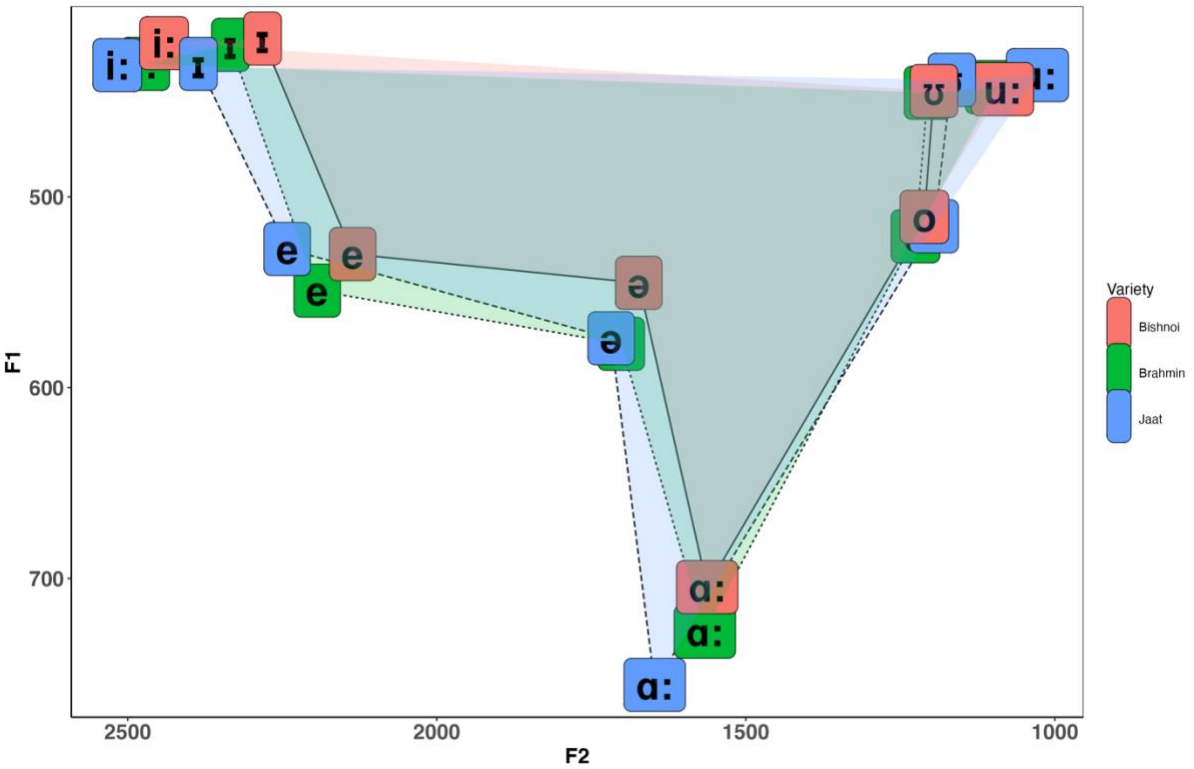


Figure 6.1 Vowel space chart created from all data types with the averages acquired from each speaker of three different varieties of Marwari based on present analysis. (Green = Brahmin, Blue = Jaat and Red = Bishnoi)

6.5.1 Spectral moments

SMs are modelled on the numerical distribution of the acoustic energy in a power spectrum of a predefined frequency range. The statistical analysis of these numerical distributions represents the four primary SMs (Nittrouer, 1995) and is often used to describe a pattern or distribution of the energy through different equations. Forrest et al. (1988) first used SMs to

analyse voiceless obstruents, and since then, they have received considerable attention in the study of other consonants, but the focus has mainly been on fricatives.

SMs have been used in assessing the place of articulation of consonants (e.g. Chistovich, 1985; Chistovich & Lublinskaya, 1979; Fox et al., 2011; Milenkovic & Forrest, 1988; Sakayori et al., 2002; Savela et al., 2007; Tahiry et al., 2016; Tuomainen et al., 2013). They have also figured in speech disorder studies (e.g. Blacklock, 2004; Colton et al., 2011; Zharkova, 2016), and child language development studies (e.g. Czaplicki et al., 2016; Körkkö, 2015; Nittrouer, 1995).

SMA has been used for forensic speaker comparison studies (e.g. Eriksson, Cepeda, Rodman, McAllister, et al., 2004; Rodman et al., 2002), but other than Weingartová and Volín (2013), the focus has been limited to consonants or long-term distribution of SMs. Weingartová and Volín (2013) emphasise the importance of examining short-term spectra of individual vowels in forensic contexts. They suggest that short-term spectra are better suited to the needs of forensic speaker comparison casework, as they obviate the requirement of having long stretches of speech for reliable extraction – forensic recordings are often of limited duration. Weingartová and Volín (2013) also suggest that using shorter slices for the analysis model means we do not have to deal with conflated category data, e.g., taking voiced and voiceless regions together. With these considerations in mind, the present study examines *spectral chunks*, i.e., short sections extracted from the middle i.e. average formants extracted from the approximate two pulses from each vowel, of the first four formants of vowels and conducts SMA for discriminating between individual speakers. The duration of these sections varied based on the extraction settings. The data was extracted from +/-3dB energy drop at either side of the peak, thus depending on this the sections were as short as 5ms and as long as 30ms.

The extraction of SMs can only be performed within a predefined frequency band. The band might be as wide as the whole range available to a human ear (20Hz – 20kHz), or any subsection (s) of that range. The present analysis was undertaken on four variable frequency bands each below 5kHz. The bands were defined by the centre frequencies of the formants under consideration.

The initial assumption here is based on the source filter theory of speech production that treats the individual vocal tract as a filter. Thus, any power spectrum would be determined by the individual speaker's 'filter,' i.e., their vocal tract shape and dimensions (Rodman et al., 2002).

6.5.1.1 First spectral moment (m_1)

The first SM is the mean energy within the spectral chunk extracted from the formant under consideration. The value is often referred to as the centre of gravity (COG) (Tuomainen et al., 2013) or centroid (Nittrouer, 1995). For clarity, we shall use the term COG throughout.

6.5.1.2 Second spectral moment (m_2)

The second SM presented by Forrest et al. (1988) is the spectrum variance, i.e., the energy variance across the spectrum. This measurement has often been substituted by the standard deviation (SD) (the square root of the spectral variance) in recent studies (Körkkö, 2015; Tuomainen et al., 2013). The SD is the deviation of the spread of power from the mean (m_1).

6.5.1.3 Third spectral moment (m_3)

The third SM, also known as the 'skewness' (Skew) or 'spectral tilt' shows the distribution of the energy on either side from the mean within the formant. A negative skewness would denote a higher concentration of energy towards the lower frequencies, and a positive skewness of the energy would show a higher energy concentration towards the higher frequencies.

6.5.1.4 Fourth spectral moment (m_4)

The fourth SM is "kurtosis" (Kurt) of a spectrum describing the peakedness of the power spectrum; thus, a lower kurtosis value suggests a relatively flat energy distribution without any clearly defined peak; conversely, a higher kurtosis indicates a peaky or narrowly pointed peak. Schindler and Draxler (2013) provided a version of the formulae to calculate SMs (p.2794):

$$m_1 = \frac{\sum a \cdot f}{\sum a} \quad (1)$$

$$m_2 = \frac{\sum a(f - m_1)^2}{\sum a} \quad (2)$$

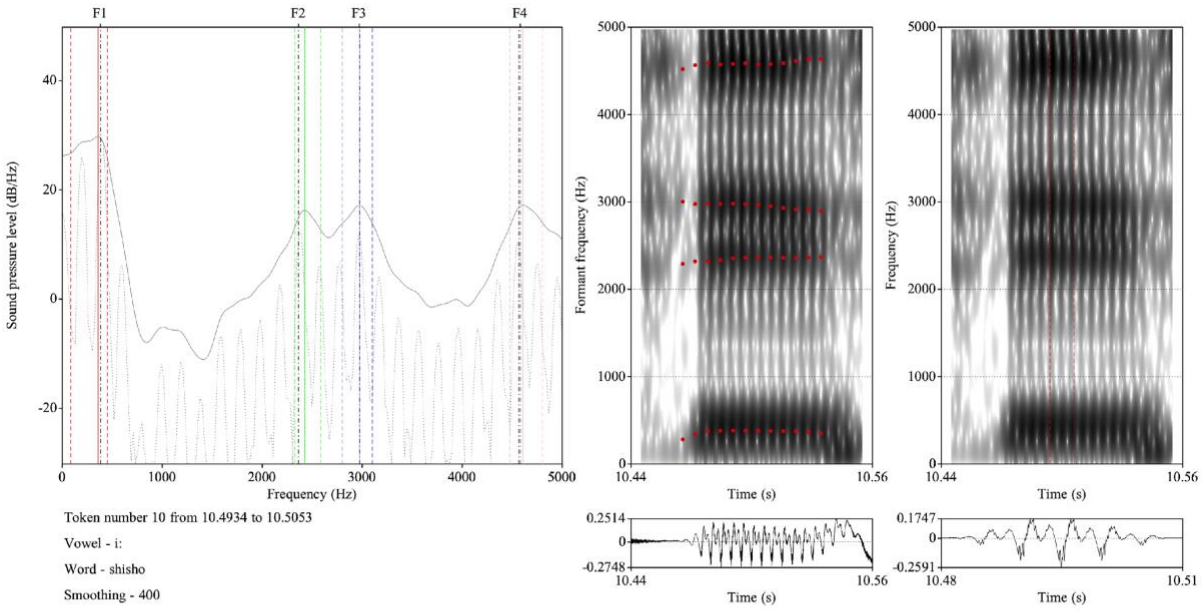
$$m_3 = \left(\frac{\sum a(f - m_1)^3}{\sum a}\right) \cdot m_2^{-1.5} \quad (3)$$

$$m_4 = \left(\left(\frac{\sum a(f - m_1)^4}{\sum a}\right) \cdot m_2^{-2}\right) - 3 \quad (4)$$

The *m* here represents the moment of the spectrum, *a* represents amplitude and *f* the frequency of the moment. Formula (1) is to calculate the COG (m_1). Formula (2) calculates the variance of the frequencies in the spectrum. To arrive at m_2, i.e., SD, one would calculate the square

root of spectral variance. Formula (3) calculates the spectrum's skewness (m_3) or spectral tilt. Formula (4) is used to calculate the kurtosis or peakedness of the spectrum (m_4). Figure 6.2 presents the SM distribution of one participant for the vowel /i:/. The figure shows how each moment was extracted and what setting was used.

As shown in the figure the settings were applied to smoothen the harmonics and make the peaks more visible (this setting was selected after multiple trials and errors). Once the script identified a sound pressure drop of 3dB on either side of the peaks, the spectral slice was extracted, and the moments were calculated (Harrison, 2021). In Figure 6.2, we can see that the peak of the F1 is skewed towards the left (lower frequencies) thus giving a negative value. Similarly, the negative kurtosis value denotes a flatter peak. COG is the mean distribution and as the multiple peaks around the first formant peak in the figure shows there were multiple energy peaks in a close proximity, providing a lower COG value from the central average value of F1.



	F1	F2	F3	F4
Lower (Hz)	86.133	2325.6	2799.3	4478.9
Centre (Hz)	358.94	2423	2972.9	4607.6
Upper (Hz)	452.2	2584	3100.8	4801.9
Width (Hz)	366.06	258.4	301.46	323
COG (Hz)	278.18	2447.7	2955.7	4637.1
SD (Hz)	93.49	62.455	72.977	79.12
Skew	-0.14789	0.16182	-0.1252	0.090705
Kurtosis	-1.1119	-1.004	-0.96193	-1.0123

Figure 6.2 Praat generated a picture depicting the 3 dB drop at either side of the formant peaks and extracted values.

While SMA has proven significant for speech articulation and synthesis studies (e.g. Blacklock, 2004; Chistovich & Lublinskaya, 1979; Colton et al., 2011; Feng et al., 2011; Milenkovic & Forrest, 1988; Zharkova, 2016), applying SMA to the four vowel formants for forensic speaker discrimination is a new approach. Substantial research has been undertaken on the speaker discriminatory value of individual formants, and some studies have examined formant bandwidths (Ishikawa & Webster, 2020; Kent & Vorperian, 2018). However, to our knowledge, no research has been undertaken on the discriminatory potential of the spectral moments of the four vowel formants.

The analysis is based on the individual vowels rather than the *isolexemic sequences* used by Rodman et al. (2002), i.e., we considered the vowels irrespective of their phonological contexts (p.26).

The research seeks to answer the following questions:

SMs as an individual speaker discriminatory measure:

1. Can an SMA of the four moments of vowel formants F1–F4 help distinguish between individual speakers?
2. Are there factors that either impede or facilitate the discriminant values of SMs? If so, what are these factors?

In that regard, we may ask,

2.1 Which SMs and Combinations of SMs are most effective?

2.2 Which vowels or subsets of vowels show the best discriminant value?

2.3 Which elicitation techniques and the associated speaking styles provide the best data for SMA?

2.4 Which varieties of Marwari does SMA work best on?

6.6 Data Processing and Analysis

Data processing started by isolating the targeted sound files for each participant. This process involved three stages: identifying the speech for analysis in individual recordings. This process included gain normalising the files in Sound Forge (9.0) by equalising the peaks to 2.0 dBFS

(decibels relative to full scale). The goal of normalizing is to adjust the amplitude of the audio signal so that the loudest peak reaches a specified level without changing the overall dynamic range. In this case, normalizing to 2.0 dBFS means that the highest peak in the audio signal will be set to a level of 2.0 decibels below full scale. Gain normalizing helps to ensure that different audio files or segments have consistent loudness levels (Maher, 2018). Normalizing is necessary for comparing or combining multiple recordings, as it prevents clipping and ensures consistent loudness across different recordings. It also simplifies tasks like feature extraction and comparison between recordings, as it sets a common reference level, ensuring a consistent listening experience.

The next step was to extract each vowel from selected words formant centre frequency values were estimated manually in Praat (6.1.54) and logged using a bespoke script (Harrison, 2021). A Gaussian window was applied, and the following settings were selected:

maximum spectrum view	'100 Hz',
pre-emphasis	6.0 d B per octave
formant ceiling	5000
Number of formants	4.5 [11]

This script logged individual formant centre frequencies up to F4 and calculated delta values (F4 minus F3, F4 mins F2, etc.). Ten tokens per vowel were extracted for each individual. At this point, it was discovered that a variety of factors, including a lack of vowel tokens between obstruents in conversational data, a higher degree of vowel reduction, and the inability to separate voices between two participants, made it difficult to process certain recordings in conversational data. As a result, the number of participants for conversational data was decreased to the best five recordings. This was distinct from the other two methods of data elicitation. Given this shortcoming, all future analyses including this technique of data elicitation will be changed accordingly.

[11] The script was configured to identify and extract 4.5 formants before automated extraction. Praat's default settings recommend a maximum formant frequency (ceiling) of 5500 Hz for women and 5000 Hz for men. By pushing the ceiling slightly higher to 4.5 formants (around 5250 Hz), the script could obtain a clearer cut off frequency for the 4th formant in female voices (Boersma & Weenink, 2001).

A further Praat script was created for identifying and extracting SMs. The script was designed to extract formants based on the manually extracted formant values, i.e., the script automatically identified the peaks closest to the previously acquired formant data and chose the nearest possible value. From these values, SMs were estimated automatically and logged.

6.7 Results

6.7.1 Impact of vowels, varieties, and mode of data elicitation on SMs

The data underwent comprehensive processing, followed by the application of linear mixed model ANOVA tests to discern the significant influence of language variety and vowels on spectral moment values. Subsequent ANOVA analyses were conducted for each mode of data elicitation, ensuring the validation of observed trends. The fixed effects considered encompassed variety, reflecting the impact of diverse language varieties on the dependent variable, vowels, signifying the influence of distinct vowel sounds, and the interaction between variety and vowel. This interaction term elucidates whether the effect of one variable is contingent on the level of the other. In acknowledging the inherent variability across participants, random effects were incorporated, introducing both random intercepts and slopes. This model structure permits the baseline level of the dependent variable to fluctuate across different participant levels, capturing nuanced variations in the data.

A full model was created to test the significance of variety and vowel (Var+Vow) as independent variables and was tested with variety alone (Var), vowel alone (Vow), and interaction between variety and vowel (Var*Vow) as part models. In summary, the analysis aims to understand the influence of language variety and vowels on SM values using a linear mixed model ANOVA approach. The presence of significant p-values indicates that these factors have a meaningful impact on the dependent variable.

The significance level for each model was tested with an ANOVA. The p-values of m_1 for every formant for these individual models were < 0.05 for all but F4 Var*Vow and F3 Var*Vow for conversation data. Table 6.1 shows p-values for each part model tested against the full model.

Table 6.1 p-values for three different models for spectral moments

Feature	Model	P-value		
		Wordlist	Story	Conversation
	Variety	<< .0001	<< .0001	0.0006
F1- m_1	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.0090	0.0689
	Variety	<< .0001	<< .0001	<< .0001
F2- m_1	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.0001	0.0002
	Variety	<< .0001	<< .0001	<< .0001
F3- m_1	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	<< .0001	0.3219
	Variety	<< .0001	<< .0001	<< .0001
F4- m_1	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.4465	0.2530	0.3718
	Variety	<< .0001	<< .0001	0.0032
F1- m_2	Vowel	0.0162	0.0116	0.0218
	Variety *Vowel	0.0003	0.0411	0.1012
	Variety	0.0062	0.0188	0.0001
F2- m_2	Vowel	0.0002	<< .0001	0.0083
	Variety *Vowel	0.2292	0.0925	0.1052
	Variety	0.008352	0.0219	0.1193
F3- m_2	Vowel	<< .0001	0.0311	<< .0001
	Variety *Vowel	0.0788	0.2092	0.3039
	Variety	0.0490	0.1250	0.7517
F4- m_2	Vowel	0.0007	0.0012	0.3024
	Variety *Vowel	0.0793	0.5968	0.5033
	Variety	0.1207	0.7766	0.5842
F1- m_3	Vowel	0.0001	0.1103	0.9447
	Variety *Vowel	0.0041	0.6868	0.5484
	Variety	0.0808	0.5492	0.2249
F2- m_3	Vowel	0.0001	0.0933	0.0075
	Variety *Vowel	0.1987	0.1265	0.3971
	Variety	0.0020	0.4072	0.8528
F3- m_3	Vowel	<< .0001	0.1767	0.0151
	Variety *Vowel	0.5675	0.3737	0.2207
	Variety	0.0703	0.9750	0.3860
F4- m_3	Vowel	0.0504	0.7573	0.3571
	Variety *Vowel	0.4749	0.0680	0.2637
	Variety	<< .0001	<< .0001	0.6458
F1- m_4	Vowel	0.0010	0.1371	0.5419
	Variety *Vowel	0.3812	0.1732	0.8478
	Variety	0.4370	0.3934	0.7509
F2- m_4	Vowel	<< .0001	0.0586	<< .0001
	Variety *Vowel	0.7775	0.1340	0.0821
	Variety	0.6642	0.6947	0.8754
F3- m_4	Vowel	<< .0001	0.0007	0.0658
	Variety *Vowel	0.1430	0.3396	0.6188
	Variety	0.1373	0.6132	0.2322
F4- m_4	Vowel	0.0025	0.2393	0.3836
	Variety *Vowel	0.0405	0.3251	0.4868

The significance for m_2 was lower than that of m_1. This implies that the significance of standard deviation was lower than centre of gravity, suggesting more variability or dispersion in the data.

The Var*Vow for the higher formants F3m_2 and F4m_2 were not significant (p-value >0.05) for all three data modes and also for F2 of story and conversation data.

The importance of variety reduced drastically for the following two SMs, with only F3m_3 and F1m_4 for wordlist data having a significant value. For m_3, none of the models had significant p-values other than wordlist Vow (F1, F2, F4) and Var*Vow (F1) for wordlist data, suggesting that the m_3 as an independent feature does not convey a lot of variety-specific or vowel-specific information.

M_4 showed significant values for all Vow wordlist data but for other data types it was only significant for the story (F3m_4) and conversation (F2m_4) data. Var model was only significant for F1m_4 (wordlist and story), suggesting m_4 alone is not a reliable feature for variety identification. Both variety and vowel were significant for the first two SMs. The interaction between them, although significant, had a lesser effect than keeping them as individual dependent variables.

The importance of variety, vowel, and their interaction for each mode of data elicitation was checked. For the wordlist context, variety, vowel and their interaction show highly significant effects ($p<0.001$). For higher moments, the effects are weaker, but some remain significant. For the conversation context, variety and vowel effects persist for F1-m_1 and F2-m_1 but interaction effects disappear. Higher moments show mostly non-significant effects. For the story context, strong variety and vowel effects reappear for F1-m_1 and F2-m_1 with some interactions. But most higher moment effects are non-significant.

To summarise, F1 and F2 moments show robust effects of variety and vowel across contexts. But higher spectral moments and interaction effects tend to weaken, especially for conversation.

6.7.2 Discrimination between individual speakers based on SMs.

Analysis of Variance (ANOVA) tests were conducted to assess the influence of variety (caste), vowel, and their interaction on spectral moment values. While the average formant values differed across caste varieties, indicating physical vocal tract differences, the impact of derived spectral moments was more minor.

Therefore, sections 6.7.2 and 6.7.3 first analysed speakers from all three caste varieties together to assess overall trends, including the role of different vowels (section 6.7.3). Individual spectral moments and their combinations were examined for each speech elicitation context (wordlists, conversation, story). Section 6.7.4 then assessed if there were any variety-specific differences in spectral moments that emerged in the combined analysis. The goal was to clarify if spectral moment patterns were driven primarily by universal factors such as vowel quality or if they also reflected more subtle physical differences across caste varieties in this dataset.

A linear discriminant analysis (LDA) was performed to evaluate how well the 16 primary spectral moments (4 moments x 4 formants) discriminate between individual speakers. LDA identifies the feature combinations that maximize separation between known speaker classes (Fisher, 1938; Martinez & Kak, 2001). It rests on assumptions of data normality and equal within-group covariances. The spectral moment data itself did not perfectly meet normality, as skewness measures will always indicate some non-normality (zero skewness indicates a symmetric distribution). However, visualization and normality tests showed the underlying formant datasets to be reasonably normally distributed before extraction of moments. To improve normality, each feature was z-normalized across speakers prior to LDA. Homogeneity of within-group covariances was tested using Levene's, Bartlett's, and Box's M tests on the transformed data. While some minor deviations from assumptions remained, LDA modelling proved reasonably robust. The goal was to assess whether spectral variation in a controlled speech context might be individually distinctive, not if datasets strictly met theoretical perfectly normal conditions unlikely in real speech. Thus LDA served as an exploratory modelling approach using spectral summaries likely to capture this variation if present.

Both correlations and covariance of the variables were tested. Figure 6.3 shows the correlations between the SMs. The correlation coefficient's magnitude indicates the strength of the linear relationship between variables. Coefficients between 0.9-1.0 suggest very high correlation, 0.7-0.9 high correlation, 0.5-0.7 moderate correlation, 0.3-0.5 low correlation, and <0.3 little/no linear correlation (Belsley, 2004).

In the figure, the size of the circles represents the magnitude of correlation, while the colour denotes the exact coefficient value of correlation. For example, the correlation between F1COG and F1 is shown by a large white circle, with the white colour indicating a coefficient value between 0.5 and 1, and the size of the circle indicating that these features are highly correlated

with one another. At the same time, the dark big circle of F3COG and F3 Skewness demonstrate that these properties are uncorrelated. The key findings from the figure are that formant frequencies are highly correlated with their respective COGs, implying that they will generate multi-collinearity concerns if included in the same model (Boedeker & Kearns, 2019).

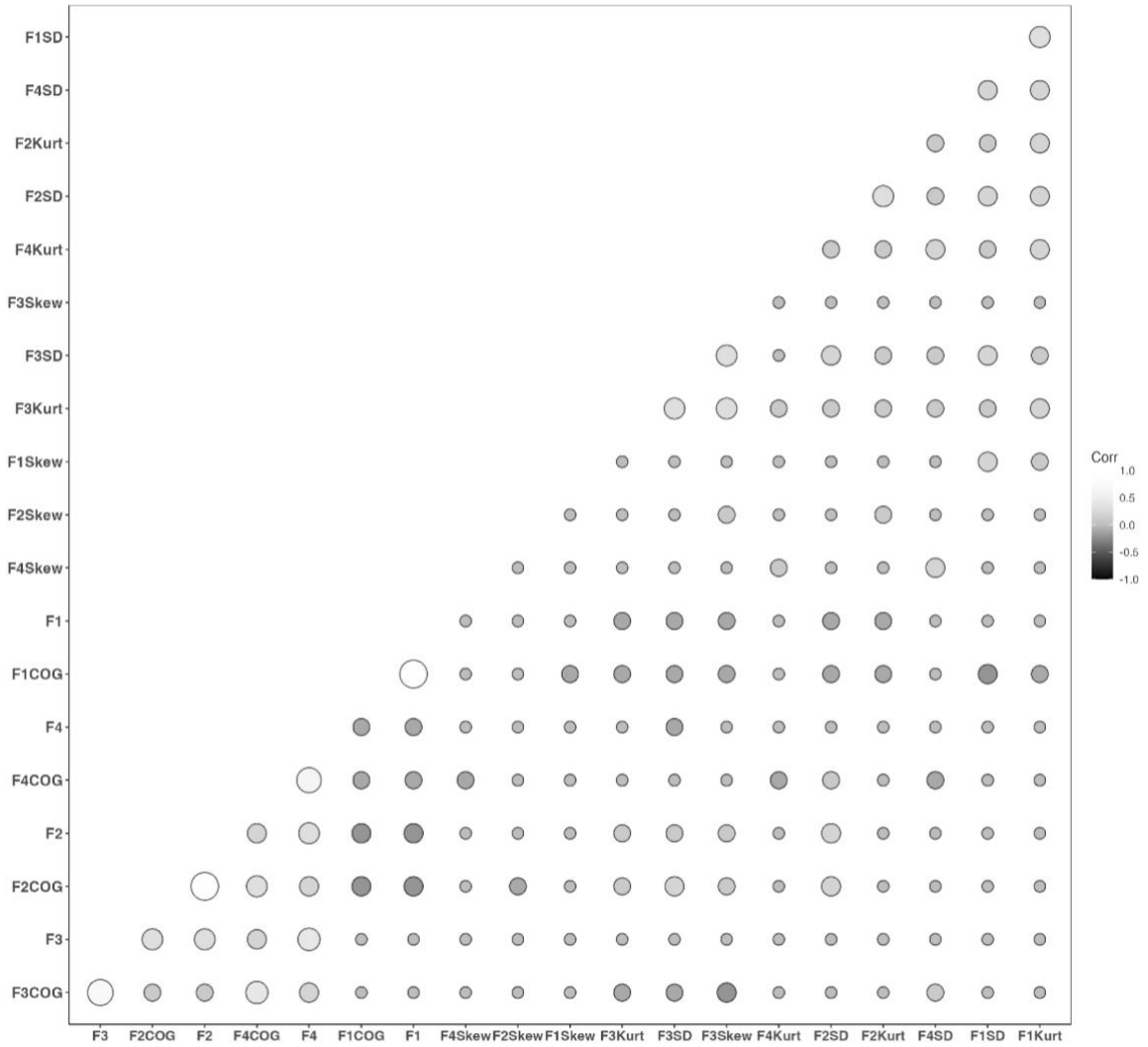


Figure 6.3 Correlation between individual SMs and centre formant frequencies. (The colour of the circle represents the range of correlation, while its size represents the degree of correlation, i.e., the larger the circle, the more correlated/uncorrelated the value.)

6.7.2.1 Speaker discriminatory power of an individual SM

The significance of SMs as individual speaker discriminants for the same variety was analysed in two steps: the speaker-discriminatory power of individual SMs and the speaker-

discriminatory power of multiple combinations of SMs. Step two included combinations of the best-performing SMs and later combining these moments with centre formant frequencies.

The analysis was conducted once all the prerequisites for an LDA were satisfied. Figure 6.4 shows the 'times greater than chance factor' of each SM. The times greater than chance level factor is shown here rather than the actual percentage because the number of participants varied due to the absence of certain feature values for some participants, and as previously stated for conversational data, this number was reduced to 5 per variety rather than 15. As a result, utilising times above chance level provided a clearer picture of how data performed in the face of shifting participant and token counts.

Figure 6.4 shows that m_1 is among the top-performing moments, surpassing chance-level classification by a factor of 5. m_2 followed closely, producing classification rates four to five times above chance for all formants. The average classification rate (CR) for wordlist and story data was 8%, nearly four times higher than the chance level. As the number of participants per variety was reduced to five for conversational data (see Section 6.6), the chance level percentage increased. Specifically, the chance of encountering conversational data rose from 1 in 45 to 1 in 15, resulting in an increase from 2.2% chance level (wordlist and story) to 6.6%. The results showed that the average CR for conversation data was 23%, and when compared to the 6.6% chance factor, analysing individual features boosted the average CR four times above chance.

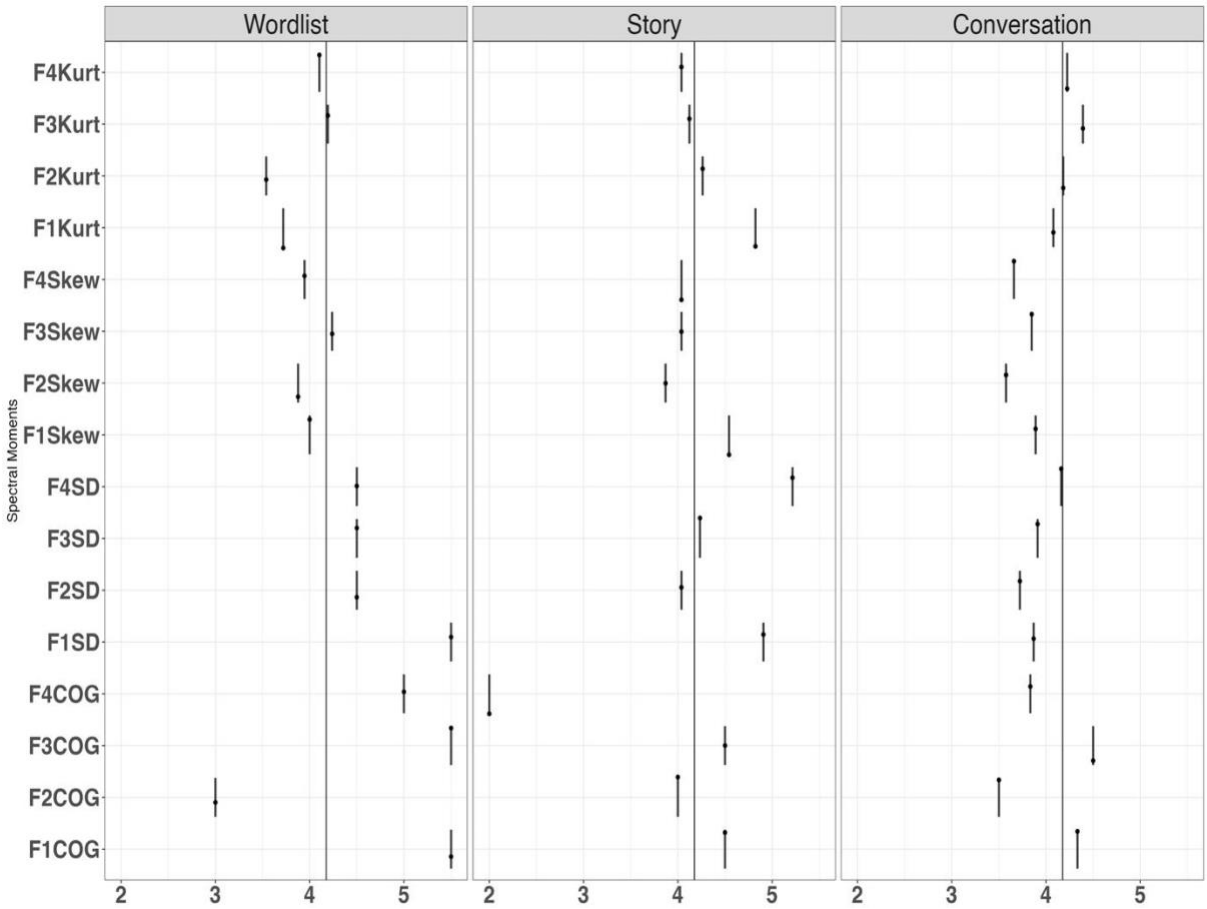


Figure 6.4 Individual CRs for three modes of data elicitation. (x-axis = classification rates over the chance level, i.e., 2-5 times above chance level; y-axis = spectral moments; vertical straight line = average CR for each mode of data elicitation)

6.7.2.2 Speaker discriminatory power of combinations of SMs

Once the individual CRs were measured, the same process was repeated for multiple combinations of SMs. This was to find a combination that provided the best CRs. The analysis was conducted in R and considered combinations of up to eight features. This number was selected because LDA only accepts *n*-1 parameters for the analysis, where *n* is the maximum number of tokens per vowel (Tabachnick & Fidell, 2007, pp. 381-382). As the maximum number of vowels extracted per person was ten, nine parameters could be used in one set of analyses. This number was further reduced to eight because some tokens were discarded during the SM extraction process.

The first step to check the significance of including SMA along with centre formant frequencies was to check if including these measurements had any impact. Table 6.2 examines the change

in CRs when spectral moments are added to centre formant frequencies ranging from one to four against the centre formant frequency values alone. The table shows that adding spectral moments does improve the CRs for speaker classification. For example, the CR using F1, F2, F3 and F4 is 8.1 for wordlist data, but this number increases when m1 or m2 is added. The table also showed that adding up to two spectral moments provides higher CRs. For example, the CR increases to 9.3 when both m_1 and m_2 are added to centre formant frequency values. However, this number start declining when more than two spectral moments are added. For example, with one exception of adding m_2 extracted from first three formants to the centre formant frequencies, every other moment dropped the CRs below 8.1 (CR acquired from centre formant frequency values alone). Overall m_1 provide higher CR rates than the other three moments and wordlist data had the highest CRs out of the three data type.

Table 6.2: Classification rates are expressed as times over chance for combinations of spectral moments and centre formant frequencies when compared to centre formant frequency alone.

Features	Wordlist	Story	Conversation
F1+F2+F3+F4	8.1	6.5	4.6
F1+F2+F3+F4+F1m1	8.4	7.1	5.0
F1+F2+F3+F4+F1m2	8.6	6.8	5.1
F1+F2+F3+F4+F1m3	7.9	6.4	4.9
F1+F2+F3+F4+F1m4	8.3	6.6	5.0
F1+F2+F3+F4+F1m1+F2m1	9.3	7.9	5.0
F1+F2+F3+F4+F1m2+F2m2	8.7	8.1	5.2
F1+F2+F3+F4+F1m3+F2m3	8.7	8.0	4.0
F1+F2+F3+F4+F1m_4+F2m_4	8.5	7.6	4.9
F1+F2+F3+F4+F1m_1+F2m_1+F3m_1	7.8	6.9	5.2
F1+F2+F3+F4+F1m_2+F2m_2+F3m_2	8.5	7.2	5.1
F1+F2+F3+F4+F1m_3+F2m_3+F3m_3	6.9	7.0	5.0
F1+F2+F3+F4+F1m_4+F2m_4+F3m_4	6.7	6.9	5.2
F1+F2+F3+F4+F1m_1+F2m_1+F3m_1+F4m_1	5.0	4.2	6.4
F1+F2+F3+F4+F1m_2+F2m_2+F3m_2+F4m_2	6.4	4.0	5.7
F1+F2+F3+F4+F1m_3+F2m_3+F3m_3+F4m_3	5.0	3.7	5.8
F1+F2+F3+F4+F1m_4+F2m_4+F3m_4+F4m_4	4.7	3.3	5.8

For each data elicitation mode, the eight best-performing SMs were selected and analysed further to see if there was any substantial increase in CR. The CR increased to 2.1 times above chance (TAC) for the conversation data, 6.08 times for the story data and 7.68 times for the wordlist data. These numbers increased drastically when centre formant frequencies were added. Table 6.2 lists the ten best feature combinations with and without centre formant frequencies.

Adding more features to the model increased the CRs drastically. Multi-collinearity issues could explain the higher performance of m_1 and formants, i.e., the CRs increased because the two independent variables were significantly correlated. Hence, a new analysis was run with models that only included the best features with less than (.6) correlation. The analysis used a correlation threshold of 0.6, meaning any pair of features with a correlation coefficient greater than or equal to 0.6 were considered moderately correlated and one of the features was removed. Typically, researchers prefer to remove highly correlated features (correlation > 0.7) to avoid issues like multicollinearity (Belsley, 2004). However, in this case, a stricter threshold of 0.6 was selected to be more aggressive in removing even moderately correlated features. This stricter approach aims to retain only the most independent and non-redundant set of features for the modelling process.

Table 6.3 shows that removing m_1 from the combinations to avoid the collinearity issue decreases the CR by fifty per cent.

Higher formants performed better on average than F1 and F2. The hierarchy of the best-performing moments could also be summarised as follows:

$$\mathbf{m_1 > m_2 > m_4 > m_3}$$

Table 6.2 Multiple combinations of best-performing moments

Features	Wordlist	Story	Conversation
Best Eight SMs			**TAC**
$F1m_1+F2m_1+F3m_1+F4m_1+F1m_2+F2m_2+F3m_2+F1m_3$	7.7	6.9	2.1
$F1m_1+F3m_1+F4m_1+F2m_2+F4m_2+F2m_4+F4m_4+F1m_4$	6.5	4.3	1.9
$F2m_4+F4m_4+F3m_4+F1m_1+F2m_1+F3m_1+F4m_1+F1m_4$	5.6	5.3	1.9
Best four SMs + Centre formant frequencies			
$F1m_1+F3m_1+F4m_1+F1m_4+F1+F2+F3+F4$	7.1	7.1	1.8
$F2m_4+F3m_4+F4m_4+F4m_1+F1+F2+F3+F4$	6.1	6.8	1.6
$F1m_1+F3m_1+F4m_1+F1m_3+F1+F2+F3+F4$	7.2	7.0	1.7
Individual SMs+ Centre formant frequencies			
$F1m_1+F2m_1+F3m_1+F4m_1+F1+F2+F3+F4$	7.4	7.2	1.9
$F1m_2+F2m_2+F3m_2+F4m_2+F1+F2+F3+F4$	9.3	7.1	2.1
$F1m_3+F2m_3+F3m_3+F4m_3+F1+F2+F3+F4$	6.1	7.2	1.7
$F1m_4+F2m_4+F3m_4+F4m_4+F1+F2+F3+F4$	6.2	6.6	1.8
			Individual SMs
$F1m_1+F2m_1+F3m_1+F4m_1$	8.4	6.3	9.7
$F1m_2+F2m_2+F3m_2+F4m_2$	5.0	4.6	4.7
$F1m_3+F2m_3+F3m_3+F4m_3$	5.0	4.2	5.2
$F1m_4+F2m_4+F3m_4+F4m_4$	3.7	4.2	4.5
F1+F2+F3+F4	8.1	6.5	4.6

Table 6.3 Combinations of best-performing moments without m_1 based on collinearity.

Features	Wordlist	Story	Conversation
Best four SMs + Centre formant frequencies			**TAC**
F1m2+F2m2+F3m3+F1m3+F1+F2+F3+F4	4.8	3.0	1.6
F4m4+F2m4+F3m4+F1m4+F1+F2+F3+F4	3.2	3.8	1.4
F1m4 +F2m2+F4m4+F2m4+F1+F2+F3+F4	4.3	3.2	1.2

6.7.3 The discriminatory power of SMs for different vowels

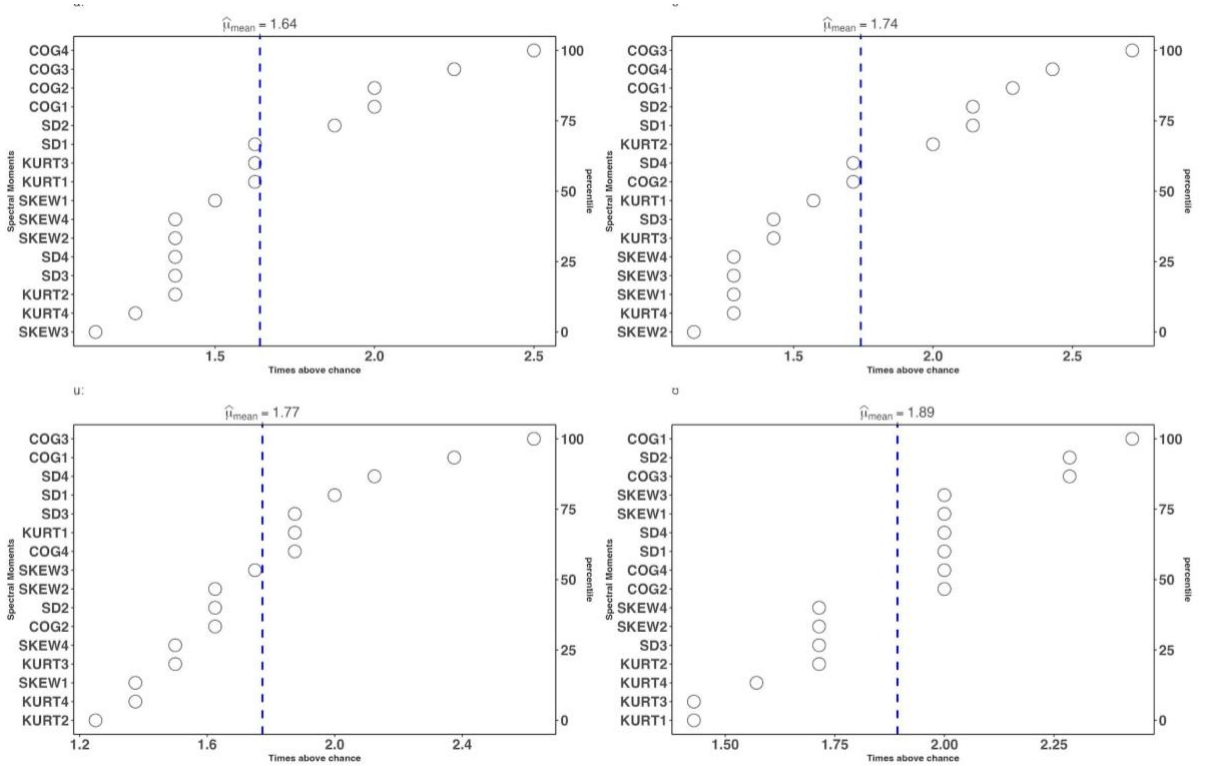


Figure 6.5 Performance of SMs for individual vowels; /a:/ (top-left), /o/ (top-right), /u:/ (bottom-left) and /ʊ/ (bottom-right) (Patil, 2021).

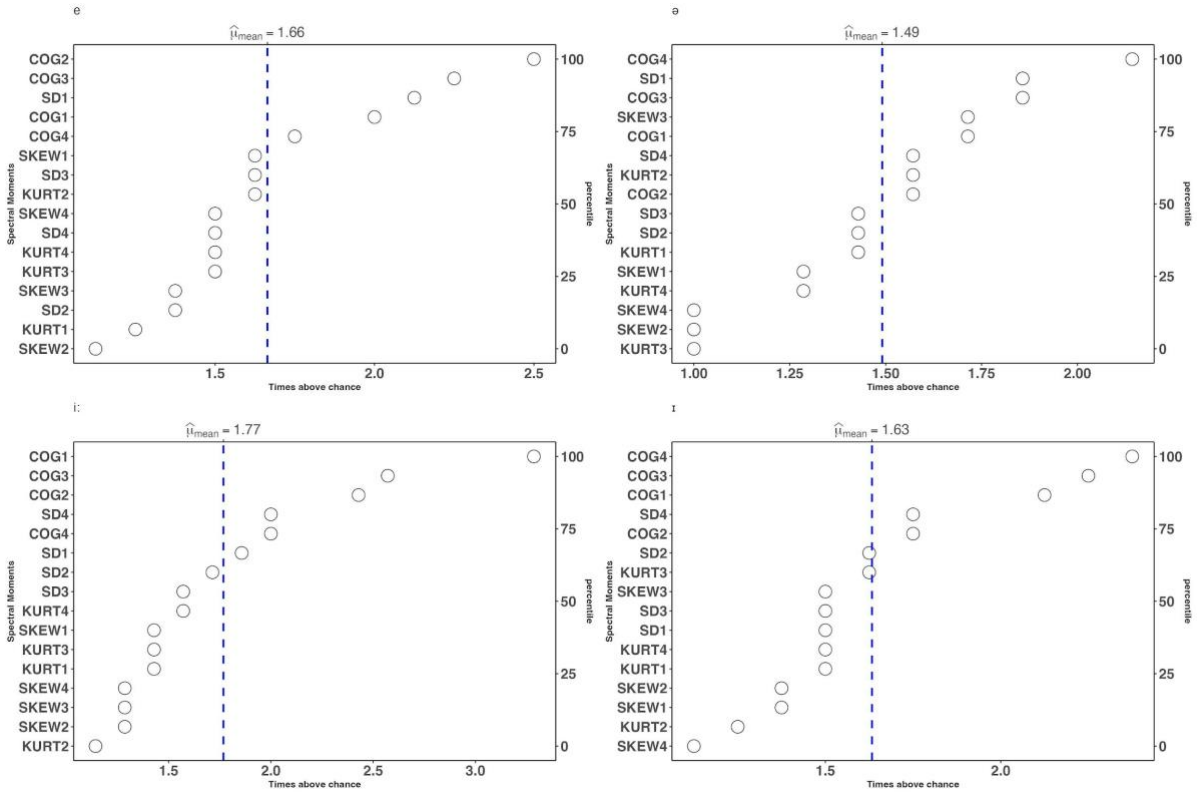


Figure 6.6 Performance of SMs for individual vowels; /e/ (top-left), /ə/ (top-right), /i:/ (bottom-left) and /ɪ/ (bottom-right) (Patil, 2021).

The effectiveness of individual vowels was also assessed. Subsetting the data by vowel further reduced the vectors for LDA. For instance, the tokens for vowel /ʊ/ reduced the LDA vectors to five due to the lack of utterances of this vowel in conversation and story data. Individual vowel analysis showed a clear pattern with back vowels /u:/ and /ʊ/ giving higher CRs than the central and front vowels. Figure 6.5 and Figure 6.6 show the results of the SMs analysis for individual vowels. The m_1 was again the best performing moment, followed by m_2. The highest means recorded were 6.38 TAC for /ʊ/. Open vowels showed a lower increase in their TAC than that of closed vowels. Within these moments, the SMs extracted from the higher formants performed better than the lower SMs extracted from the lower formants.

The performance of these vowels can be summarised as follow:

/ʊ/>/u:/>/e/>/o/> /ə/> /i:/> /ɪ/

For the wordlist data, the vowel /u:/ had the best CR; for the story and conversation, it was /ʊ/. The m_1 and m_2 combined (SM1) average had the best CR. Vowel /o/ performed very well for

wordlist, but this performance declined for the other two types. Even though vowel /a:/ performed five times above the chance level for wordlist data, overall, it performed consistently low for all three modes of data elicitation (Figure 6.7). This could be interpreted as the SMA performing relatively poorly for the open vowels; however, since there was only one open vowel included in the analysis, this possibility needs to be further tested.

Vowel subsets were further tested for combination models. Three combination models from the four best-performing SMs were selected.

$F1m_1+F2m_1+F3m_1+F4m_1$ (SM1)

$F1m_1+F3m_1+F4m_1+F2m_2$ (SM2)

$F2m_4+F4m_4+F3m_4+F1m_1$ (SM3)

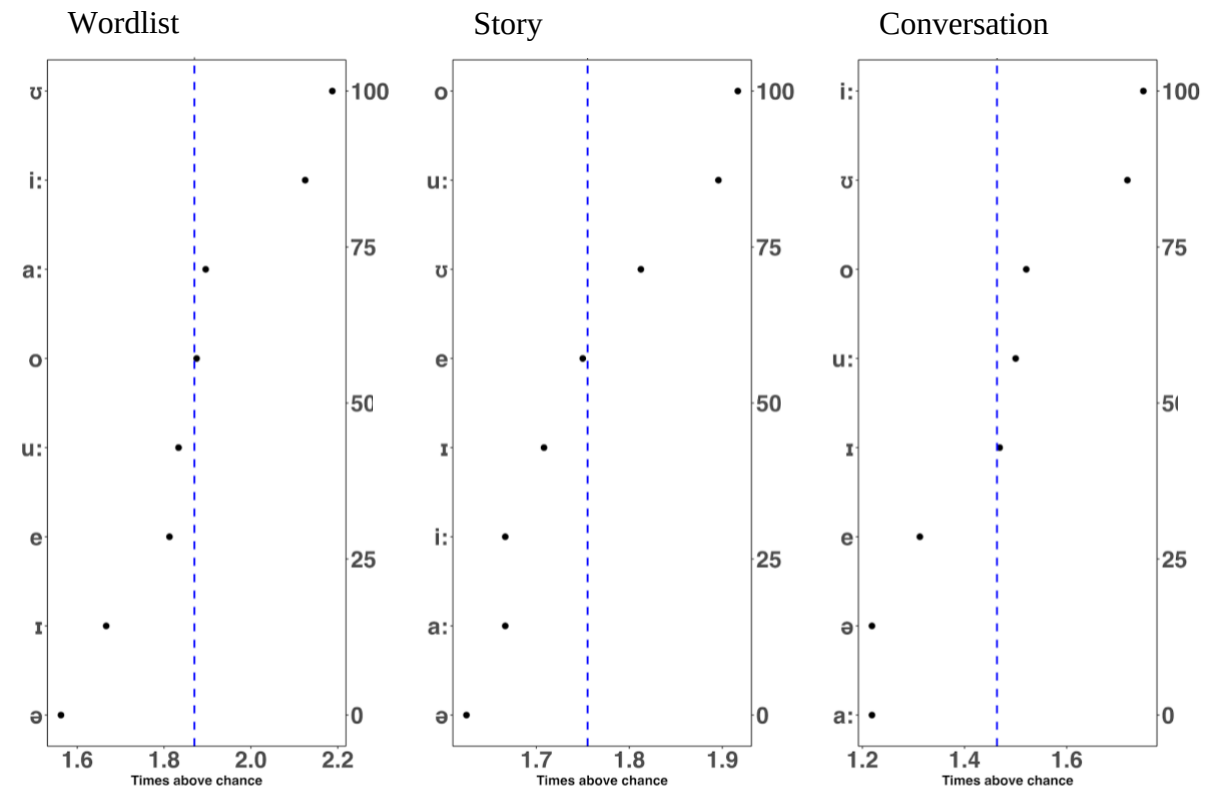


Figure 6.7 Performance of vowels for three different modes of data elicitation (Patil, 2021).

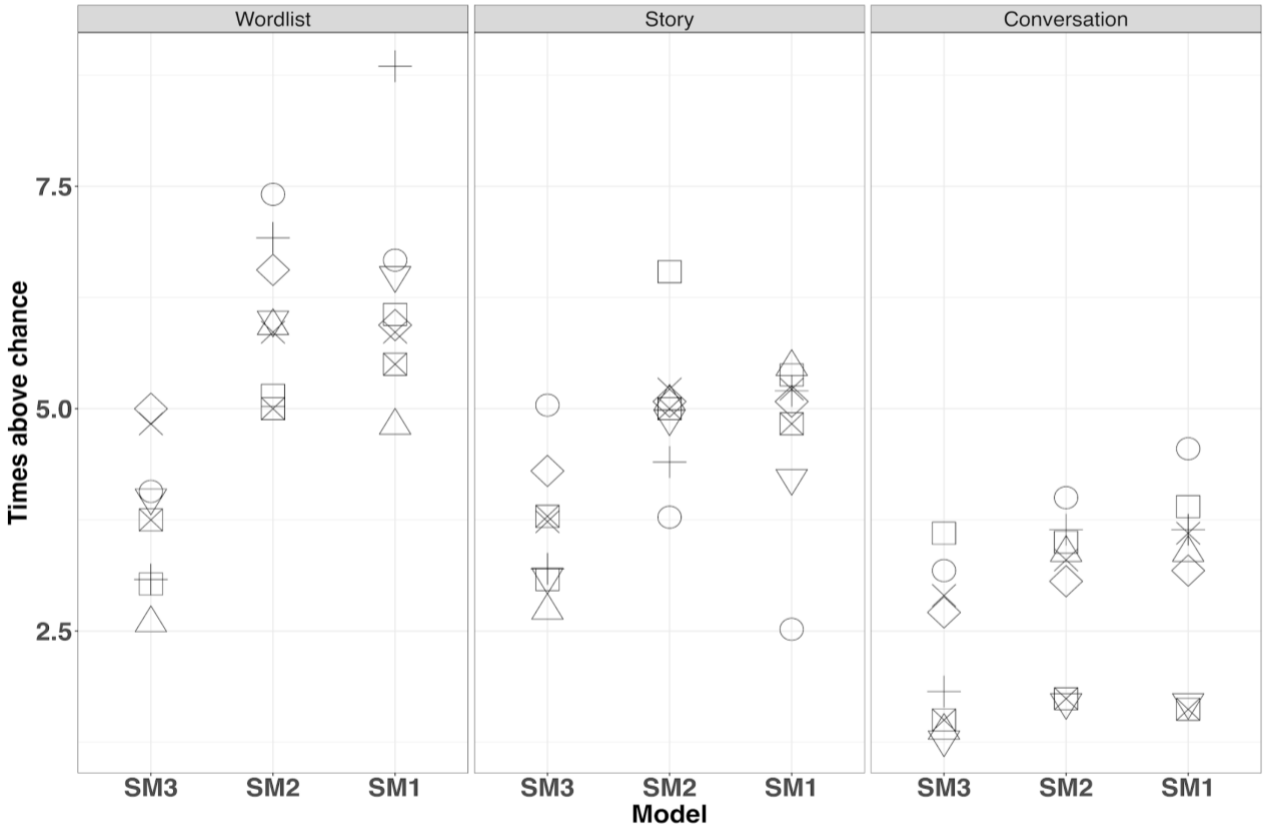


Figure 6.8 Performance of three SM models for different vowels and data elicitation types.

SM1[F1m_1+F2m_1+F3m_1+F4m_1] had the best outcome for every vowel in general, thus suggesting the importance of m_1 in speaker classification studies (Figure 6.8).

6.7.4 The discriminatory power of SMs for different varieties

Two separate models were compared to assess the effects of SMs on different varieties. The first model (M1) consisted of the SM CRs for all varieties together for every vowel and the second model (M2) was a subset of these vowels for each variety (Figure 6.9).

The performance of SMs decreased drastically once they were compared for individual varieties due to the reduction in the number of participants, thus reducing the number of tokens for the model to train on. The next step was to determine how individual SM performed for each variety. This step was divided into two parts: the performance of SM or combinations of SMs on different varieties for every vowel together, and the same step was repeated for individual vowel subsets. Each model was also tested for different data elicitation types. The results are presented in Table 6.5 and 6.5.

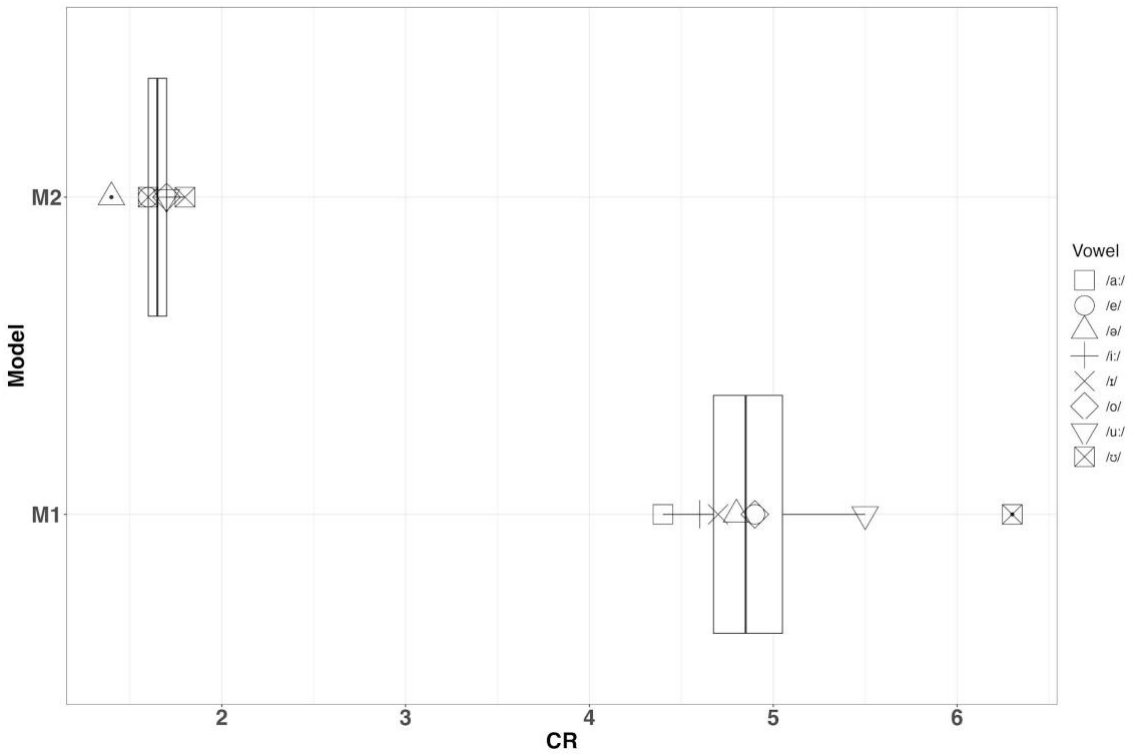


Figure 6.9 Difference between CRs for M1 and M2.

Table 6.4 shows if including SMs can improve the CRs of centre formant frequencies. This analysis was conducted for each variety subset, thus with a smaller dataset. The highlighted rows in the table represents the collinear features. As presented in the table, although the collinear features did increase the CR rates, but that increase was also observed for non-collinear features. The table also supports the initial hypothesis that including SMA along with centre formant frequencies will increase the accuracy of the system. Including the same spectral moment for each formant along with centre formant values raised the times above chance level of CRs from 1.7 to 2.2 (m_1), 2.3 (m_2), 2.4 (m_3), and 2.3 (m_4), which is at least 220 per cent increase from the chance level.

Table 6.4 Performance of centre formant frequencies and SMs

	Brahmin			Jaat			Bishnoi		
	Wordlist	Story	Conversation	Wordlist	Story	Conversation	Wordlist	Story	Conversation
F1+F2+F3+F4	1.7	1.8	1.1	2.1	1.5	1.5	1.8	1.6	0.9
F1+F2+F3+F4+F1m_1	1.9	2.1	1.3	2.2	1.2	1.7	1.3	1.8	1.0
F1+F2+F3+F4+F1m_2	1.8	2.1	1.4	2.2	1.6	1.5	1.7	1.8	0.8
F1+F2+F3+F4+F1m_3	1.7	1.8	1.1	2.1	1.3	1.4	1.5	1.7	0.9
F1+F2+F3+F4+F1m_4	1.7	2.1	1.0	2.0	1.4	1.4	1.7	1.8	0.8
F1+F2+F3+F4+F1m_1+F2m_1	2.2	2.5	1.3	2.4	1.7	1.6	2.1	2.3	0.7
F1+F2+F3+F4+F1m_2+F2m_2	1.8	2.2	1.4	2.2	1.7	1.7	2.2	2.0	0.8
F1+F2+F3+F4+F1m_3+F2m_3	1.9	2.0	1.2	2.2	1.5	1.5	2.1	2.0	0.7
F1+F2+F3+F4+F1m_4+F2m_4	2.1	2.2	1.3	2.0	1.5	1.6	1.9	1.9	0.8
F1+F2+F3+F4+F1m_1+F2m_1+F3m_1	2.4	2.4	1.5	3.5	1.8	1.7	2.2	2.3	0.7
F1+F2+F3+F4+F1m_2+F2m_2+F3m_2	2.2	2.4	1.5	3.8	1.7	1.8	2.7	2.6	0.7
F1+F2+F3+F4+F1m_3+F2m_3+F3m_3	2.0	2.2	1.1	3.0	1.5	1.5	2.2	2.6	0.6
F1+F2+F3+F4+F1m_4+F2m_4+F3m_4	2.0	2.6	1.3	4.4	1.6	2.0	2.0	2.5	0.8
F1+F2+F3+F4+F1m_1+F2m_1+F3m_1+F4m_1	2.2	3.3	1.4	3.9	1.9	1.7	3.2	3.3	0.5
F1+F2+F3+F4+F1m_2+F2m_2+F3m_2+F4m_2	2.3	3.5	1.7	3.9	1.9	2.4	2.9	3.4	0.7
F1+F2+F3+F4+F1m_3+F2m_3+F3m_3+F4m_3	2.4	3.3	1.6	3.4	1.7	1.9	3.2	3.2	0.6
F1+F2+F3+F4+F1m_4+F2m_4+F3m_4+F4m_4	2.3	3.2	1.9	4.6	1.7	2.5	2.8	3.6	0.7

Table 6.5 Best-performing SMs across all vowels for different varieties

	Brahmin				Jaat				Bishnoi			
	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean
TAC	2.8	2.3	2.6	2.5	2.0	2.6	1.6	2.06	3.3	2.6	2.3	2.7
Moment	F3m_1	F3m_1	F1m_2		F3m_1	F1m_2	F3m_1		F1m_1	F2m_1	F3m_1	

Table 6.6 Performance of combination of SMs across all vowels

	Brahmin				Jaat				Bishnoi			
	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean
SM1	4.5	4.3	1.9	3.5	2.6	2.3	2.5	2.4	3.1	3.5	1.9	2.8
SM2	4.8	4.7	2.1	3.8	2.8	3.0	2.4	2.7	3.5	3.3	1.7	2.8
SM3	2.1	2.6	1.5	2.1	1.8	1.4	1.4	1.5	2.2	2.1	1.0	1.7
Mean	3.8	3.8	1.8		2.4	2.3	2.1		2.9	2.9	1.5	

Table 6.5 and 6.6 present the cumulative results of different SMs and the three models created earlier from best-performing SMs. For individual SMs, these results show that m_1 again performed better than the other SMs. Although the average individual variety differences were 2.5 (Brahmin), 2.06 (Jaat) and 2.7 (Bishnoi), in general, these measures provided similar increases above the chance levels for each one of these. This suggests that the significance of SMA for variety differentiation is consistent and could be utilised in the future.

Table 6.6 reports the differences between the three models, which showed that SM3 provided the minimum increase in the CRs (which is different from the results acquired in section 6.7.3, where variety subsets were not created), while SM2 was overall the best-performing feature combination model. Between the different modes of data elicitations, conversation was again the worst-performing mode, while wordlist provided the highest CRs.

Table 6.7 Performance of individual SMs for different vowels for each variety and data type presented with times above chance values and the best-performing spectral moment.

	Brahmin				Jaat				Bishnoi			
	Wordlist	**Story**	**Conversation**	**Mean**	**Wordlist**	**Story**	**Conversation**	**Mean**	**Wordlist**	**Story**	**Conversation**	**Mean**
/aː/	3.0 $F1m_1$	3.0 $F4m_1$	2.0 $F2m_1$	2.6	3.0 $F1m_1$	2.0 $F1m_1$	2.0 $F2m_1$	2.3	4.0 $F4m_1$	3.0 $F4m_1$	--	3.5
/e/	4.0 $F2m_1$	3.0 $F1m_1$	2.0 $F1m_1$	3.0	3.0 $F1m_2$	3.0 $F3m_1$	2.0 $F1m_1$	2.6	4.0 $F2m_1$	2.0 $F3m_1$	--	3.0
/ə/	3.0 $F3m_1$	3.0 $F3m_3$	2.0 $F2m_1$	2.6	--	3.0 $F4m_1$	3.0 $F2m_1$	3.0	2.0 $F1m_1$	2.0 $F2m_1$	--	2.0
/iː/	4.0 $F1m_1$	4.0 $F1m_1$	3.0 $F1m_1$	3.6	--	3.0 $F3m_1$	3.0 $F1m_1$	3.0	--	3.0 $F1m_1$	4.0 $F1m_1$	3.5
/ɪ/	3.0 $F4m_1$	3.0 $F1m_1$	2.0 $F3m_1$	2.6	2.0 $F1m_1$	3.0 $F4m_1$	3.0 $F4m_1$	2.6	3.0 $F1m_1$	4.0 $F3m_1$	--	3.5
/o/	4.0 $F1m_1$	4.0 $F3m_1$	2.0 $F4m_1$	3.3	--	3.0 $F3m_1$	3.0 $F4m_1$	3.0	--	3.0 $F3m_1$	3.0 $F1m_1$	3.0
/uː/	5.0 $F3m_1$	4.0 $F1m_4$	2.0 $F1m_1$	3.6	3.0 $F3m_1$	3.0 $F1m_1$	2.0 $F1m_1$	2.6	3.0 $F3m_1$	3.0 $F1m_1$	--	3.0
/ʊ/	--	3.0 $F2m_2$	3.0 $F1m_1$	3.0	3.0 $F1m_1$	2.0 $F1m_1$	5.0 $F3m_3$	3.3	3.0 $F1m_1$	--	--	3.0
Mean	3.7	3.3	2.3		2.8	2.7	2.8		3.1	2.8	4.0	

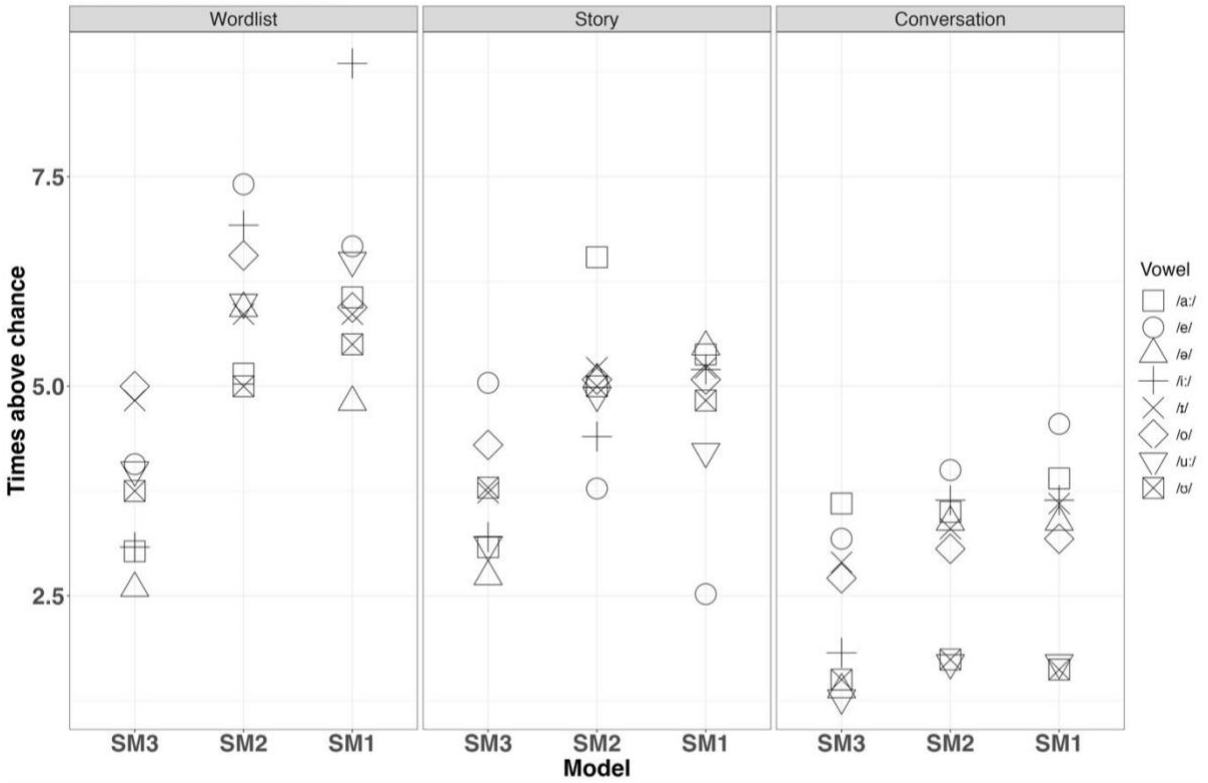


Figure 6.10 Performance of different models for individual varieties for every vowel.

Different subsets of vowels were created for each variety and analysed for each mode of data elicitation (Table 6.7)[12]. The general trend suggested a steady best-performing SM, i.e., m_1. The highest CR was observed for close vowels. The mean CRs varied from as high as 3.6 to as low as twice chance level. There were some vowels where the model could not be tested because of the lack of an n-1 number of tokens for that variety/vowel. For different modes of data elicitations, wordlist had the highest average rates for Brahmin. For Jaat, both wordlist and conversation showed an increase of 2.8 times above chance. Although for Bishnoi, the highest TAC level value was presented by the conversation data, there were only two vowels that had enough tokens to be analysed, thus making this result less dependable.

The final part of the analysis was conducted on the different vowel subsets for the three models created in section 6.7.3. The results showed that Brahmin had the highest CRs above the chance level, especially for the wordlist data. Figure 6.10 also shows the averages of these results (represented as a line for each mode of data elicitation). The highest averages were recorded for wordlist (Brahmin) and conversation (Bishnoi). For each of these, the average performances reached up to 5.5 times higher than chance level, suggesting a high increase from the single SM values. SM1 and SM2 performed better than SM1, which is congruent with the results from

Table 6.6, where the subsets of the vowels were not created. These results suggest that SM1 and SM2 are better-performing models for speaker classification.

6.8 Summary and Discussion

We began with a set of research questions, which for ease of reference, are reproduced here:

Q1. Can a SMA of the four moments of vowel formants F1–F4 help distinguish between individual speakers?

A1. The data here clearly shows that there is value in including SMA in the battery of the speaker discrimination tests (Table 6.4).

[12] The empty spaces in the table are for the vowels/features for which n-1 number of tokens were not available. The loss of these tokens could be the result of error in spectral moment extractions or limited number occurrence for that particular vowel.

Q2. Are there factors that either impede or facilitate the discriminant values of SMs? If so, what are these factors?

A2. The results vary depending on how the technique is applied and to which material it is applied. The specific factors investigated here were found to impact the efficacy of SMA.

So,

Q2.1 Which SMs and Combinations of SMs are most effective?

A2.1 Centre of gravity (m_1).

Q2.2 Which vowels or subsets of vowels show the best discriminant value?

A2.2 SMA works better when applied to certain vowels and vowel combinations. However, there is no clear pattern here as the vowels/combinations depend on speaking style.

Q2.3 Which elicitation techniques and associated speaking styles provide the best data for SMA?

A2.3 SMA works best for wordlist data, followed by story and conversation.

Q2.4 Which varieties of Marwari does SMA work best on?

A2.4 There was no clear distinction between the performance of SMA for all three varieties, thus suggesting that SMA works for all three varieties equally.

6.9 Explanations of the Findings

We now turn to some putative explanations of our findings:

Q1/A1. The first question is why SMA should be a good speaker discriminant at all. It is well established that formant frequencies – commonly measured as centre values – are not only linguistically determined, i.e., by the vowel system of the language and variety spoken but are also biologically affected by the geometry and configuration of the resonating chambers of an individual speaker's vocal tract. Given that this is so, we would regard SMA as being a refinement of simple centre frequency measurements, and, therefore, capable of carrying more nuanced or detailed information about the tract. Acoustically, spectral moments offer a deeper understanding of vocal tract formant energy distribution compared to just analysing centre

formant frequency values. For example, COG values reveal whether a speaker's energy is more concentrated around the lower regions (formants) or higher regions.

However, in the present state of knowledge, we are not able to link specific moments to specific aspects of the biological tract. This situation of first determining whether a particular analysis process is a good discriminator before being able to motivate it in terms of individual speaker physiology has precedents in speech science, notably in respect of MFCCs, which have been seen to be an extremely high-performing individual speaker discriminant, even though we remain unable to specify which particular coefficient might correspond to which particular component(s) of the tract. As with MFCCs, further research is needed into the interface between voice and speech production settings and processes and their acoustic reflexes.

Q2.1/A2.1. **The best performing SM** is m_1; although calculated differently, this is very similar to the traditional measures, i.e., formant centre frequency. It is not surprising that given the established pedigree of formant centre frequencies in discrimination tasks, it should perform equally, if not slightly better than the centre frequency. This hypothesis was supported by the findings of this study, as both m_1 and formant centre frequencies increased the CR 8 times above the chance level. Given that this is so, it is not surprising that adding more detailed within-formant spectral information should enhance its discriminatory value. Nor is it surprising that including m_1 values from all formants should outperform simply including it for one formant or a limited number. For fricatives, m_1 is inversely proportional to that of the oral cavity in front of the point of constriction (Jongman et al., 2000). Even though fricatives and vowel spectra cannot be compared to each other, as the first one carries a flatter spectral distribution throughout the spectrum than the second one and vowels have decay on higher frequencies, the direct correlation between m_1 and formants could provide some new insights on the relationship between m_1 and the oral cavity.

Regarding m_3 and m_4, the least well-performing moments, the data presented here can shed little light on their relative reduced performance. One possibility which would need to be investigated in any follow-up research concerns is the possibility of their being unstable and prone to high intra-speaker variation. Given that the best speaker discriminants are those which show the greatest inter-speaker variability and least intra-speaker variability, any feature that fails on the later count may contribute little to the test. However, individual m_2 from the mean m_1 is one measure that may turn the need for intra-speaker variation stability on its head and

marshal the lack of it to discriminatory advantage- i.e., wide degrees of variations for any individual speaker might mark them out from the mainstream.

The initial hypothesis that SMs extracted from higher formants of vowels might have significant implications for speaker discrimination was proven to be accurate to some extent, as although F3 SMs performed significantly better the majority of cases this was followed by F1 SMs.

Table 6.2 shows the further interaction of SMs with centre formant frequencies (other than m_1) for combinations of moments with a significant four-times increase above the chance level. This suggests that SMA could be considered one of the critical measures for speaker discriminant and classification studies.

Q2.2/A2.2. addressed two different questions:

If the vowels selected for analysis are impacting on the SMA? The first question was addressed in section 6.7.1, while looking at the mode of data elicitation, vowels showed a predominantly significant impact on the p-values for the wordlist dataset. These results could be explained by the fact that the wordlist data did not contain any unstressed vowels, consequently providing more robust values for the analysis. The results are also congruent with the findings of Themistocleous et al. (2016), which presented a significant difference between the measurements of fricative followed by stress vs unstressed vowels. Since story and conversation data had both stressed and unstressed vowels, the same rationale can also be applied here. Tahiry et al. (2016) worked on the SMA of Arabic vowels and showed that vowel SMA could be divided into two phases, a transient phase (at the beginning of the vowel) and a steady phase (when the vowel stabilises). Their work also indicated that the steady phase of vowels is least susceptible to changes. The current moments were extracted from the steady phase of the vowel, thus making the measurements more robust for speaker and vowel classification. m_1 and m_2 were again the best-performing moments, followed by m_4.

For question two, **which vowels or subsets of vowels show the best discriminants of SMs**, results showed that vowels significantly impacted SMA, thus further strengthening Savela et al. (2007)'s claims that the perceptual similarities between vowels could be interpreted by adding SMA and analysing formants. There was a definite pattern forming with vowels throughout the investigation, i.e., even if some vowels such as /i:/, /ɪ/ performed poorly when analysed in isolation, for SM models, the same vowels coupled with /o/, /u:/, /ʊ/, and /e/

consistently performed better than the rest. This could suggest that close vowels, which tend to have closer formant peaks, have an advantage of SMA over formant analysis. This could be because close vowels, produced with the tongue closer to the palate, tend to have more distinct and well-separated formant peaks compared to open vowels (Stevens, 1989). This clearer formant structure, results in a more consistent and stable acoustic signature for closed vowels across different contexts. Consequently, SMA may be better equipped to accurately capture and model the acoustic patterns of closed vowels, while formant analysis alone could struggle more with the less well-defined formants and increased contextual variations found in open vowels. However, both of these explanations are theoretical, and further research is required to have a clearer understanding of why close vowels may perform better than open vowels.

Furthermore, there was no clear pattern regarding which vowel performs best as a classifying variable for their SMs. As presented in Figure 6.8, even though the CRs for some vowels were as high as nine times higher than the chance (e.g., /i:/), this was different for each mode of data elicitation. This could suggest that the SMs of vowels are affected by the mode of speech and, as a result, affects the performance of the LDA.

As mentioned earlier, stressed, and unstressed vowels tend to have different SMA results; thus, story and conversation data could have behaved differently from the wordlist. M_3 has been shown to be significant in inter-speaker variability for Czech Vowels (Volín & Zimmermann, 2011; Weingartová & Volín, 2013). Within different vowels also, some vowels performed better at discriminating speakers than others. Weingartová and Volín (2013) showed that closed vowels presented the most inter-speaker variation (/i/ and /u/) while studying Czech vowels. The current results concur with these findings for wordlist data but not for story and conversation data. For these two modes, there was no clear pattern observed. Though high-back vowels did show that they perform better than the rest, the results need to replicate on bigger sample sizes and lab-acquired recordings.

Q2.3/A2.3. whether the **mode of data elicitation significantly impacts SMA.** The present study's results did not agree with the previously established work that shows no significant differences between different modes of data for obstruents (e.g., Kardach et al., 2002) . As most studies on SMA for modes of data elicitation focused on consonants rather than vowels, this research significantly contributes to the field. The initial ANOVAs to determine vowels and a variety-specific influence were all influenced by the mode of data elicitation. wordlist data

showed more promising results, followed by story and then conversation. This suggests the importance of the data elicitation mode for any SMA. There has been some work on how different speaking styles might influence the performance of SMA. Sadiq and Harwardt (2011) showed the difference between normal vs loud voice SMA. In general, most of the work has been focused on the SM difference rather than the impact of this difference on speaker classification. Eriksson, Cepeda, Rodman, Sullivan, et al. (2004) tested the effect of imitation on speaker comparison tasks, but other than this study, there was no reference available. This makes the findings from the present paper more noteworthy. Even though wordlist was the best-performing data elicitation mode, there was a minimal difference between story and wordlist CRs. For a combination of best-performing SMs, story CRs were even better in some instances. Though both data elicitation modes are forensically unrealistic, the story data was collected in a narration manner rather than read out loud. This could benefit future forensic case works where any incident narrated recordings are being analysed.

Q2.4/A2.4. Exploring **the discriminatory power of spectral moments (SMs) across different varieties** of the same language, the ANOVA analysis suggested some variety-specific differences for the first two moments (m_1 and m_2). The Jaat variety performed worse than the other two, which could be attributed to the inter-variety vowel space differences. Although these results were minimal and did not indicate significant inter-variety disparities, the vowel space chart in Figure 6.1 revealed that Jaat has more open and fronted vowels compared to the other varieties, making it more distinct. However, this apparent distinctiveness in vowel space did not translate into substantial differences in the extracted spectral moments derived from the formant values. Despite the varieties occupying different vowel spaces, denoting differences in average formant centre values, the spectral moments calculated from these formants did not exhibit significant variety-specific variations. These findings could suggest that factors beyond just vowel space distinctions, such as speech rate, recording conditions, and the need for more participants or tokens, may need to be considered to draw clearer conclusions about the discriminatory power of spectral moments across language varieties. While the vowel spaces differed, the spectral moment features extracted from the formant values were not significantly impacted by these variety-level distinctions.

For fricative studies, these SMs have been associated with representing the place of articulation (m_1-m_4) and voicing (m_1 and m_3) (Jongman et al., 2000). However, unlike fricatives, the energy concentration for the higher frequency ranges starts declining for vowels, thus resulting in less

robust results for all four moments. The non-significant results for F4 m_1 and m_2 can also be linked to the reduction of energy distribution in the higher formants of the vowels. Although the current work only focused on variety-specific differences, the results varied from those proposed by Eriksson, Cepeda, Rodman, Sullivan, et al. (2004) for Swedish and English vowels. They applied the MER (minimal enclosing rectangle) method to the two languages, showing that SMs are language-independent (Eriksson, Cepeda, Rodman, McAllister, et al., 2004). These differences can be explained by the non-dynamic formant extraction process utilised by the present study, where instead of using the SMs from a larger *chunk* or *isolexemes* as suggested by Eriksson, Cepeda, Rodman, Sullivan, et al. (2004), the SMs were extracted from the vowel formant mid-points. This kind of extraction would be more susceptible to the proceeding and following phonemes, thus providing different results from the previous studies.

On the other hand, the dynamic analysis suggested by Eriksson, Cepeda, Rodman, Sullivan, et al. (2004) also makes the lexemes more vulnerable to different speech processes. This could be eliminated by the present study's vowel-midpoint extraction. The variety-specific differences noted in the present study could also help with a more concrete acoustic differentiation between various dialects, further strengthening the claim proposed by Clopper and Pisoni (2004), which states that auditory judgements of regional dialects from listeners have always been dependent on the local knowledge of the investigator.

6.10 Limitations

One limitation of this study is that the number of participants was limited, especially for the conversation data. This, combined with the number of tokens uttered per participant, limited the number of variables that can be put into the same LDA model. The model needs to be tested with more participants and tokens uttered per participant to further assess the significance of SMs. Another limitation of the present study is that the process of manually extracting SMs is very lengthy and time-consuming. The current procedure though effective, needs to combine with the already present ASR-based mechanism, to be more efficient and time-saving. The third limitation would be that only one language was used for the present study. The same model needs to be tested on other languages, especially the ones that are linguistically further away from Marwari language. This would show the efficiency of this model for different languages and language families. The impact of speech rate, phonological environment and forensically realistic data also needs to be tested.

6.11 Implications

The results for 45 speakers and their SMA indicate that including SMA in forensic casework might significantly increase the correct CR of an individual speaker. The study also suggests that SMs extracted from the steady state of vowels (Tahiry et al., 2016) could help with forensic research, especially if the extractions were made from stressed vowels.

Although further testing is needed to assess the significance of SMAs for different settings of data elicitation, it is recommended to include SMA in addition to formants, F0 and voice quality analysis for any manual speaker discrimination analysis.

7 Chapter 7 Presented as Article 2

7.1 Research Degree Thesis Statement of Authorship

University of York

York Graduate Research School

Candidate name	Nikita Suthar
Department	Language and Linguistic Science
Thesis title	Within-formant spectral feature analysis for forensic speaker discrimination casework: A study of 45 Marwari monolinguals from Bikaner, India

Title of the work (paper/chapter)	Within-formant spectral measures and their role as a source of speaker discriminant information	
Publication status	**Published**	
	Accepted for publication	
	Submitted for publication	*
	Unpublished and unsubmitted	
Citation details (if applicable)	NA	

Description of the candidate's contribution to the work	Conceptualisation, literature review, data collection and analysis, writing and manuscript preparation, citation, and references
Percentage contribution of the candidate to	90%

the work	
Signature of the candidate	Nikita Suthar
Date (DD/MM/YY)	25th September 2023

Co-author contributions*

By signing this Statement of Authorship, each co-author agrees that:

(i) the candidate has accurately represented their contribution to the work;

(ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co-author	Prof. Peter French
Contact details of co-author	peter.french@york.ac.uk
Description of the co-author's contribution to the work**	Part of conceptualising, contribution to interpretation of findings, editing
Percentage contribution of the co-author to the work	10%
Signature of the co-author	
Date (DD/MM/YY)	27th Sept 2023

7.2 Title: Within-Formant Spectral Measures and Their Role as a Source of Speaker Discriminant Information

7.3 Abstract

Formant analysis has been used as one of several methods for speaker discriminant studies (Cao & Dellwo, 2019; McDougall, 2006; McDougall & Nolan, 2007). Most studies have focused only on formant centre frequencies, trajectories (McDougall & Nolan, 2007), and to a more limited extent, bandwidths (Fleischer et al., 2015; Gonzalez-Rodriguez, 2011; Kent & Vorperian, 2018). The current study takes the potential role of formants as individual speaker discriminants further by investigating a range of within-formant measures. It reports on work conducted on the amplitude, relative amplitude, spectral bandwidth, LPC bandwidth and spectral peaks of formants. Marwari language was used as a testbed, and in principle, this analysis could be conducted on any other language. Marwari belongs to the Indo-Aryan language family and is spoken in Rajasthan, India. A total of forty-five female Marwari monolingual speakers from the Bikaner district were recruited for the study. Recordings were collected from spontaneous and non-spontaneous speech and focused on eight vowel phonemes. Three modes of data collection were employed. The first mode of data collection was a list of 80 words (10 tokens per vowel) that the participants were asked to read aloud. The second mode was a picture description task, i.e., participants were shown a picture of local deities and were asked to narrate a story associated with the deity. The third method was a conversation where participants were paired and asked to have an unscripted conversation on a topic of their choice or chosen from a provided list. An ANOVA conducted in R showed an impact of vowel and variety on every mode of data elicitation. Once these differences were established, the goal was to examine individual speaker discrimination. Spectral measures were extracted from the first four formants. Manually assisted and corrected automatic formant extractions were conducted using a Praat script (Harrison, 2021). Then a linear discriminant analysis (LDA) was conducted on the measures extracted from every formant to predict the classification rate of these measures in identifying individual participants. The results show that front vowels provide more information than back vowels for speaker classification and that some measures (amplitude and spectral peaks) perform exceptionally better than others.

Keywords: Forensic speaker comparison, vowel formants, spectral measure analysis, amplitude, spectral peak, spectral bandwidth, LPC bandwidth

7.4 Introduction

Linguists have advocated for the manual articulatory-acoustic analysis of speech samples for forensic speaker comparison since it was first used in the late twentieth century. Along with the articulatory aspect, these methods also look at the various acoustic elements of the speech sample, including the spectral analysis of vowels and consonants. The analysis of vowel formants and their importance in forensic speaker comparison work has been the focus of multiple studies (e.g.,Cao & Dellwo, 2019; Fleischer et al., 2015; Gonzalez-Rodriguez, 2011; Kent & Vorperian, 2018; McDougall, 2006; McDougall & Nolan, 2007; Nolan & Grigoras, 2005) While it is known that lower formants (F1 and F2) are better suited for vowel identification, higher formant frequencies such as F3 (e.g., McDougall, 2004) and F4-F5 (e.g., Cao & Dellwo, 2019) of vowels have been suggested to carry speaker-specific information.

The focus of forensic speaker comparison work has recently shifted from manual acoustic or articulatory analysis to automatic speaker recognition-based systems (ASRs), and it has become one of the prime methods of speaker comparison and discrimination for forensic casework13 (French, 2017; Gold & French, 2019; Hughes et al., 2018). ASR is also used commercially for voice comparison and verification cases (Watt et al., 2020).

Acoustic features tested in comparison to ASR, such as formants (Patil et al., 2010), amplitude (Mitra et al., 2012), bandwidths (Gonzalez-Rodriguez, 2011), SMA (Eriksson, Cepeda, Rodman, McAllister, et al., 2004) etc., have also shown potential for forensic casework (Weingartová & Volín, 2013).

The present study looks at the three acoustic measures mentioned above; formant amplitude, formant bandwidth, and spectral peak of vowel formants to assess their significance in speaker comparison works. These measures were extracted from the first four vowel formants (F1-F4).

[13] United Kingdom government has not yet admitted ASR for speaker identification (see French (2017)).

7.4.1 Formants

Many researchers have argued for vowel formant centre frequencies as a parameter to be used in forensic speaker comparison casework (Cao & Dellwo, 2019; Jessen, 2008; McDougall, 2004). Formants are considered as being determined or constrained by the geometry of an individual speaker's vocal tract (Peterson & Barney, 1952). Research has demonstrated the impact of including higher formants - particularly F3 - (e.g., McDougall, 2004) or F4 and F5 (e.g., Cao & Dellwo, 2019) in speaker discrimination and comparison casework. The first two formants are useful in determining to which vowel phoneme an uttered vowel might belong, but higher formants are more helpful for identifying individual speakers.

A lot of forensic case material is based on recorded telephone calls. Therefore, most studies on higher formants have focused on F3 alone since the bandwidth of telephone communication is generally limited to 3500 Hz. However, with recent advances in the underlying technology of social media or mobile communications, band limitations set by telephonic conversations have been improved. Due to these technological advances, formants beyond F3 can now be extracted from certain recorded telephone conversations (Cao & Dellwo, 2019).

The present paper hypothesises that including formant amplitude, bandwidth and spectral peaks in speaker comparison work will improve manual acoustic analysis methods.

7.4.2 Formant amplitude

Amplitude is defined by Ladefoged (1996) as "*the maximum variation in air pressure from normal (p.16).*" Measuring the maximum amplitude of any complex wave is not a straightforward process (Ladefoged, 1996). There are multiple methods of extracting amplitudes (including formant amplitude), such as spectrum envelop (A_s), root mean square (A_e):, average amplitude (A_a), initial voice period peak (A_i), and peak amplitude (A_p) (Fant & Artony, 1963). One such method of amplitude measurement that can easily be extracted with the help of a Praat script was selected; sound pressure level extracted at the nearest maximum of a formant.

Amplitude variations for different phonation types of a speech signal can be perceived by lay listeners as loudness (Lindblom et al., 2009). Lindblom et al. (2009) also mention that while a lay listener can perceive amplitude and other measures, understanding a vowel without the presence of these measures is also possible. However, the perception of formant amplitude

remains controversial for vowel perception studies (Kiefte et al., 2010). Recent studies have demonstrated that listeners struggle to identify vowel sounds when only provided with formant frequency information (Bladon & Lindblom, 1981; Ito et al., 2001; Miller, 1984). To better understand vowel perception, researchers must consider the broader spectral properties of speech rather than just isolated acoustic cues. Keeping this mind, the role of these global spectral measures, especially amplitude still needs to be checked for inter-speaker variations.

While analysing amplitude, relative amplitude (RA) is often also used for spectral analysis as it *determines the spectral balance* between two formants (Aaltonen, 1985, p. 2). RA has been calculated in different ways depending on the research goals. Some studies looked at the relative amplitude values of harmonics while trying to understand breathiness (Fischer-Jørgensen, 1968; Hillenbrand et al., 1994). Others have looked at the RAs of different formants while trying to understand how the human ear perceives vowel quality (e.g., Aaltonen, 1985; e.g., Ainsworth & Millar, 1972; Carlson et al., 1970; Lindqvist-Gauffin & Pauli, 1968; Miller, 1953). The present study follows the second method of extracting RA values, where the difference between two formant amplitude values have been analysed.

7.4.3 Formant bandwidth

Formant bandwidth refers to the width of peaks (formants) in the frequency spectrum of speech sounds like vowels (Fleischer et al., 2015). It specifically measures the frequency range around a peak where energy is significant, quantified as the width at -3 dB from the peak (Lindblom & Sundberg, 2014). Narrower bandwidths indicate more precise vocal tract resonances and articulation, while wider bandwidths suggest less precision (Hawks & Miller, 1995). For example, research on dysphonia has revealed wider bandwidths in some types of disordered speech (Fleischer et al., 2015). Overall, bandwidth provides information on articulation precision and vocal tract filtering (Fant, 1972). Lindblom and Sundberg (2014) defined it as acoustic energy loss from multiple factors like radiation and viscosity. They showed an open glottis markedly increased first formant bandwidth. In the frequency domain, they defined it as the width 3dB down from the formant peak, implying larger bandwidths produce flatter peaks. Recent work by Schwartz et al. (2018) demonstrated extensive speaker-specific information in higher formant bandwidths. Considering this, the current work analyses bandwidths for the first four formants.

Bandwidths, unlike amplitudes, are difficult measures to extract. The process of extracting accurate formant bandwidth measurements can be challenging. This is because bandwidths fluctuate within each pitch period between the closed and open phases of vocal fold vibration (Medabalimi et al., 2014). In other words, the bandwidth is not constant but changes dynamically with the glottal cycle. During glottal closure when the vocal folds are together, formants are more precisely articulated, and bandwidths are narrower. During glottal opening when the folds separate, formants become less distinct, and bandwidths increase. This intra-cycle bandwidth fluctuation relates to changing vocal tract shape and filtering over the glottal cycle. In practice, this means bandwidth estimates can vary depending on which part of the cycle is measured. Overall, the intrinsic linkage between glottal cycle events and formant bandwidth makes accurate extraction challenging. This issue is often solved by using short speech segments for the extraction. Various other methods of bandwidth extractions have been proposed by different studies, such as AM-FM modelling (Cohen et al., 1992), estimating bandwidth modulations with the help of instantaneous frequency signals (Medabalimi et al., 2014), exponentially weighted autoregressive (EWAR) spectral models (Potamianos & Maragos, 1995), bandwidth estimation from decaying constants of resonance frequencies (Yasojima et al., 2006), extracting bandwidth from group delay function (Medabalimi et al., 2014), and linear predictive (LP) analysis (Makhoul, 1975; Reddy & Swamy, 1984). In the present study, the two methods used are those in Praat (Boersma & Weenink, 2001), i.e., spectral bandwidth (SB) and linear predictive coding or LPC bandwidth (LB).

Formant bandwidth cues have been associated with the identification of the sex of the speaker, since analysing bandwidth for female voices provides results very different from those of males. Females have higher bandwidths (Kent & Read, 2002). One proposed explanation for higher female formant bandwidths comes from Hanson and Chuang (1999). Their research suggested females may have a greater tendency for posterior glottal openings or chinks compared to males. These glottal gaps could contribute to the wider bandwidths observed in female speakers.

In addition to sex identification, formant bandwidth also plays a role in vowel perception and quality, though its impact is more nuanced. Vowel perception is rarely affected by their respective bandwidths, but any drastic reduction or increase in the bandwidth can make vowels sound artificial (Hawks & Miller, 1995; Kent & Read, 2002). A series of seminal studies have shown that varying formant bandwidth has minimal impact on vowel identification, even with

extremely narrow or zero bandwidths (Stevens et al., 1969; Remez et al., 1981). While bandwidth alterations do not affect perception, they significantly influence naturalness, with abnormally narrow bandwidths sounding distinctly artificial (House, 1960). Intelligible speech can even be synthesized using just three sinusoids at formant frequencies, demonstrating that precise bandwidth cues are not essential for comprehension if listeners expect speech (Remez et al., 1981). However, as bandwidth increases, vowel distinctiveness decreases due to overlapping formants (Fant, 1970). Nasalization demonstrates this through diminished formant peaks and reduced vowel differentiation (Lindblom et al., 1977). Thus, while bandwidth does not appear critical for identification, there may be optimal values for maximizing clarity and naturalness.

Considering formant bandwidth's role in optimizing clarity and naturalness, though not acting as an isolated cue, the current study will examine incorporating bandwidth metrics into a manual model for speaker discrimination. Despite bandwidth not directly identifying speakers, including its quality-enhancing attributes could potentially improve model performance. This paper will assess whether complementing the representation with bandwidth factors that are not cues themselves, but shape perception can strengthen speaker discrimination ability. The goal is to capitalise on bandwidth's impacts on quality to augment an identification model reliant on spectral features.

7.4.4 Spectral peak

Spectral peaks are the peaks extracted from the nearest maximum of estimated formants. They are often impacted by the level of cepstral smoothing for the spectrum. Poles, formants, and resonances together can be categorised as spectral peaks (Rossing, 2014). Once the sound signal had been processed through an LPC filter, the harmonic smoothing was applied for the spectral peak extractions. In most cases, spectral peaks coincide with formant peaks, but the results are impacted by smoothing parameters.

Lindblom et al. (2009) discuss the significance of further spectral features which may help a listener identify minute characteristics of individual speech sounds. As mentioned earlier, formant amplitude (Kiefte et al., 2010), formant bandwidth (Klatt & Klatt, 1990) and spectral peaks (Hillenbrand & Houde, 1995) are often considered perceptual measures for vowel identification studies. The present study is focusing on these perceptual measures only. It tries to assess if an acoustic analysis of these features can complement forensic casework.

While individual studies have analysed the potential of bandwidth (Ishikawa & Webster, 2020; Kent & Vorperian, 2018) and amplitude (Ainsworth & Millar, 1972; Kiefte et al., 2010) for forensic speaker comparisons, the present study is the first to test these features in combination. Moreover, this study is the first to consider spectral feature analysis of F4 for the purpose of speaker discrimination.

7.5 Data

The present study was conducted on the Marwari[14] language. It focuses on the acoustic differences between three different caste-based varieties of Marwari, i.e., Brahmin, Bishnoi and Jaat. Traditionally, Brahmin, the highest Varna (occupation-based hierarchical system of Hindus), were priests. Jaat belongs to the 'Vaisya'[15] Varna, the working or business population of the caste system. This caste was predominantly involved with farming and herding cattle in Rajasthan. The Bishnoi caste or community is the newest of the three castes, created by Jambheshwar Ji (in 1485) to overcome the caste system. Initially, people who wanted to avoid the caste system started following twenty (bi:s) - nine (noi) rules and created a community called "Bishnoi." Over time, the community has begun identifying as a separate caste. Most members of this caste work in agriculture as well. To exclude regional variation as a variable, the recordings were collected from long-term resident female monolingual speakers from the Bikaner district, who have never left the district in their lifetimes.

The collected speech material showed that the three caste-based varieties exhibit some phonetic differences. As presented in Figure 7.1, the vowel space of the Jaat variety has more fronted and open vowels compared with the other two. Eight different vowels were selected because of their presence in each selected variety. The eight vowels selected here were:

[iː], [ɪ], [e], [ə], [aː], [o], [uː], [ʊ]

[14] Marwari is an Indo-Aryan language spoken mainly by the members of the Marwari community (also called Marvari, Marvadi and Marwadi) residing in the north-western areas of Rajasthan (a state in northwest India).

[15] Vaisyas are the biggest population group in Varna system, consisting of Carpenters, goldsmiths, ironsmiths etc.

Forty-five participants were recruited, with 15 speakers for each variety. All participants were born and raised in the Bikaner district. They are all aged above 40, with a mean age of 50.68 years (range 40-84, standard deviation = 8.03). Each participant was asked about their educational qualifications and linguistic competence. The primary criterion for selecting these participants was their monolingualism. All participants were naïve to the specific research questions but understood the general purpose of the study. Speakers from the Bishnoi and Jaat varieties predominantly reside in the rural areas of the district. Brahmins, on the other hand, live in urban areas. The first author visited the participants' houses and made the recordings in the quietest non-echoic room.

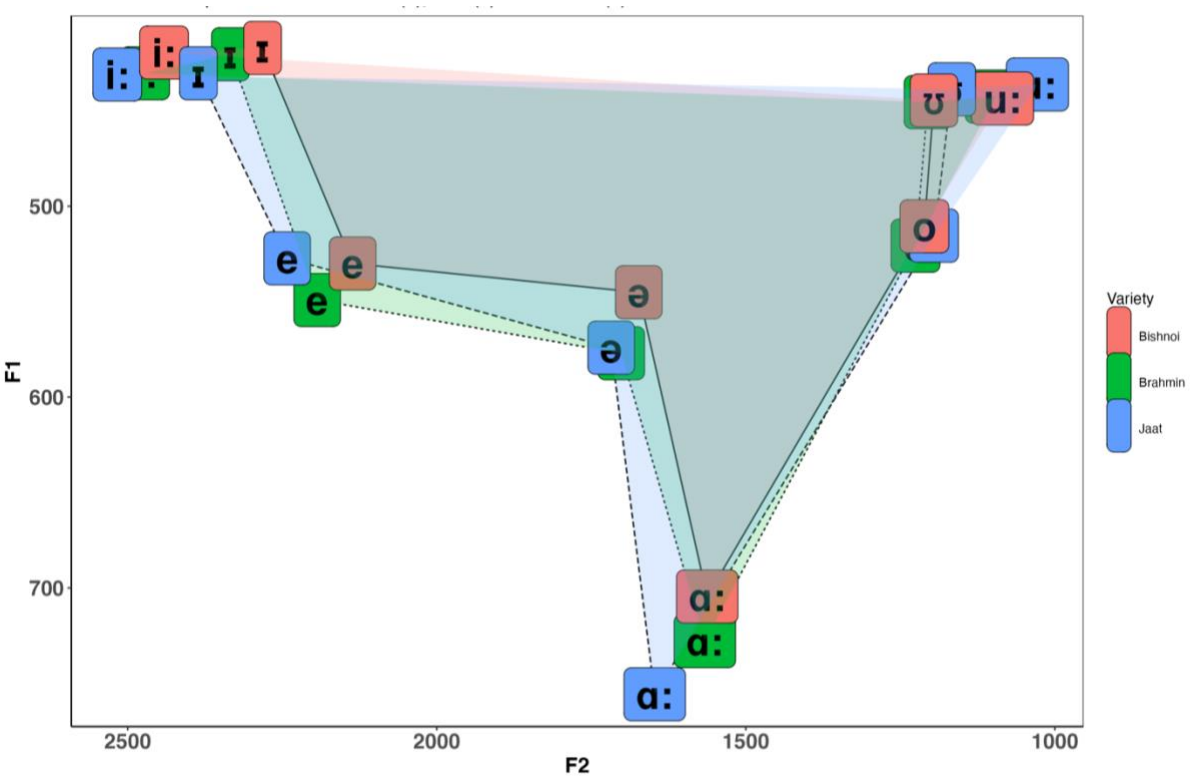


Figure 7.1 Vowel space chart of three different varieties of Marwari created from all three types of data (Suthar & French, 2023a). (Green = Brahmin, Blue = Jaat, and Red = Bishnoi)

The recordings were collected from both spontaneous and non-spontaneous speech. Three modes of data collection were employed. The first mode was wordlist, where the study adapted Swadesh's (1955) methodology to select culturally appropriate terms with target vowel positions (Ladefoged, 2003; Adi-Bensaid & Tobin, 2010). However, low female literacy rates necessitated using Ladefoged's (2003) community pronunciation technique where a reader (informant from the same variety as the participant) articulated words for participants to repeat.

Although practical, this community pronunciation risks convergence, obstructing idiolectal details (Pardo et al., 2022). Initial assumptions were high convergence given interlocutor age and gender differences (Babel, 2009; Earnshaw, 2021). However, participants asserted their forms were correct when divergences arose, suggesting more divergence than expected (Giles, 1973). This divergence could stem from the researcher's presence and inquiries about differences (Pardo et al., 2018). Ultimately, methodology did not negatively impact results, demonstrating that for unwritten dialects, adapted approaches may sufficiently capture phonemic distinctions. Careful consideration of participant demographics and goals is warranted when balancing practical techniques against ideal models for gathering comparable lexical data across language varieties.

The second mode was a picture description task, i.e., participants were shown a picture of local deities and asked to narrate a story associated with them.[16] The third method was a conversation where two participants were paired and asked to either discuss a topic of their choice, or a topic from a list provided. All three modes of data collection were conducted in one sitting for each informant to avoid variations in emotional, biological, or physical stress. Participants were recorded with a high-quality digital recording device, 'Zoom H4n Handy Recorder[17]' (files: .wav format; 44.1 kHz sampling rate; 16-bit depth). This recorder came with two built-in microphones that could be adjusted to 90 degrees or 120 degrees, as required. The recorder was positioned 25 centimetres from the participant's mouth on a tripod to create the same environment for every participant. The recorder's microphones were adjusted to 120 degrees for both channels depending on the participant's position.

7.6 Research Questions

The goal of the present paper is based on the hypothesis that including within-formant analysis in speaker discrimination work can contribute to a forensic phonetician's test battery. The article seeks to answer the following questions:

16 This task came natural to the participants as religion is deeply intertwined with the culture of the region, and since the Marwari community has a strong oral tradition.

17 Specifications: https://www.zoom.co.jp/sites/default/files/products/downloads/pdfs/E_H4nSP_0.pdf

RQ1) Are spectral measure values impacted by variety, vowel and mode of data elicitation?

RQ2) Can including spectral measures with formant centre frequencies help distinguish between individual speakers in an acoustic analysis?

If yes,

RQ2.1) Which spectral measures and combinations of spectral measures are most effective?

RQ3) Are there any factors that impede or facilitate spectral measures' discriminant values?

If so, we may ask,

RQ3.1) Which vowels or subsets of vowels yield the highest classification rate (CR) results when spectral measures are applied?

RQ3.2) Which speech styles provide the highest CRs when spectral measures are applied to them?

RQ3.3) Which varieties do spectral measure analysis work best on?

7.7 Data Processing and Analysis

Data processing started by isolating the targeted sound files for each participant. This process involved three stages: Identifying the required sound sections (words) from individual recordings, processing them by removing any section with noisy background, and analysing them in Sound forge (9.0) and Praat (1.8.3). The process included gain normalising the files. This procedure, conducted in Sound forge, equalised the peaks to 2.0 dBFS. Any sound file that still contained background noise after the process was discarded.

The next step was to extract each vowel from the target words, and formant centre frequency values were logged using a Praat script. The following settings were used for the extraction:

Window shape:	Gaussian
Maximum spectrum view:	100 Hz
Pre-emphasis:	6.0 dB
Method of spectrum analysis:	Fourier
Formant ceiling:	5000 Hz
Formants:	up to 4.5
Dynamic range:	30 dB

This script logged individual formant frequencies up to F4 of each extracted vowel and their differences. Each file was double-checked for any errors and manual corrections were applied where needed.

A new Praat script was created for the subsequent analysis step, i.e., identifying and extracting spectral measures (Harrison, 2021). The script first smoothened the harmonics to make the formants more visible, making it easier to pick them up automatically. The script was designed to extract formants based on the manually extracted formant values, i.e., the script automatically identified the peaks closest to the previously acquired formant data and chose the nearest possible values. Formant amplitude was extracted with the help of the script. It extracted the peak amplitude based on the sound pressure level of the nearest formant peak for individual formants (A1-A4 for F1-F4). The relative amplitudes (RA) between two formant peaks were also analysed, as this is one of the most used methods for amplitude analyses. The amplitude difference between the two formants provided six different RA measures: Amplitude of F2-Amplitude of F1 (**A2-A1**), Amplitude of F3-Amplitude of F1 (**A3-A1**), Amplitude of F4-Amplitude of F1 (**A4-A1**), Amplitude of F3-Amplitude of F2 (**A3-A2**), Amplitude of F4-Amplitude of F2 (**A4-A2**) and Amplitude of F4-Amplitude of F3 (**A4-A3**).

Spectral bandwidth was extracted by calculating the difference between +/- 3dB upper and lower frequency for formant peaks.

$$f_nSpecBW = f_nUpperFreq - f_nLowerFreq$$

$$(f_nLowerFreq = f_nSP - (f_nAnalysis\ band/2)$$

$$f_nUpperFreq = f_nSP + (f_nAnalysis\ band/2))$$[18]

The script also extracted the LPC bandwidth for further analysis. Spectral Peaks (SPs) were extracted from the nearest maximum of the manually extracted formants.

18 The analysis bands selected for smoothing (filtering windows) were modified by trial and error to find the optimum settings. The final bands selected with the least number of errors were F1 = 300, F2 = 500, F3 = 600, F4 = 700.

7.8 Results

7.8.1 Effects of vowels, varieties, and mode of data elicitation of a language on spectral measures

RQ1 asked if spectral measure values were significantly affected by vowels, variety or mode of data elicitation. Linear mixed model ANOVA (analysis of variance) testing was conducted on the spectral measures to check if the values were affected by the variety or vowels. The analysis was repeated for data elicitation mode to corroborate the results. One full model with variety and vowel (Var+Vow) as independent variables, and three-part models with variety (Var), vowel (Vow), and interaction between variety and vowel (Var*Vow) were created to assess the impact of these variables. A linear mixed model ANOVA was conducted for each one of these part models with the full model to test the significance level of each variable. Figure 7.2 presents the p-values of each part model (x-axis) for every spectral measure (y-axis).

As represented in Figure 7.2, for wordlist data, all three models performed significantly better (the p-values were less than 0.05) for most spectral measures with very few exceptions. The significant p-values indicated that evaluating these part models against the full models with the specified variable (vowel alone, variety alone, or their interaction) had a substantial influence on the extracted values of the selected features (amplitude, bandwidth and spectral peaks). Most insignificant p-values were obtained by models that contained variety as an independent variable, both when tested alone or in combination with vowel (LB1, SB2, SB3, A4), suggesting that for wordlist data, the null hypothesis (variety did not have any impact on spectral measure values) was not rejected.

The mode of data elicitation had a significant impact on the spectral measures. As represented in Figure 7.2, the number of non-significant p-values (>0.05) increased for story and conversation, which shows that the mode elicitation affects the values of spectral measures. While the rise in insignificant p-values is suggestive, direct statistical testing is necessary to definitely confirm an influence of elicitation mode on spectral metrics, as well as describe the type and magnitude of that effect. This effect will be later tested with the help of linear discriminant analysis in Section 7.8.7.

Overall, wordlist and story data produced some insignificant p-values for the two models that assessed variety and the interaction of variety and vowels. In this connection, the story data

produced a lower number of significant p-values than wordlist data. Conversation data, on the other hand, led to drastically higher number of insignificant p-values for these two models.

Both models suggest that neither variety nor the interaction of variety with vowels did affect the extraction of spectral measures. The increased number of less significant p-values for conversation data suggests that the extracted values did not significantly differ for different varieties.

Overall, the 'vowels alone' model presented the greatest number of significant p-values in determining the values of spectral measures for all three modes of data elicitation, and the other two models had fewer significant p-values. Once this had been determined, the next step was to assess the role of spectral measure for discriminating speakers.

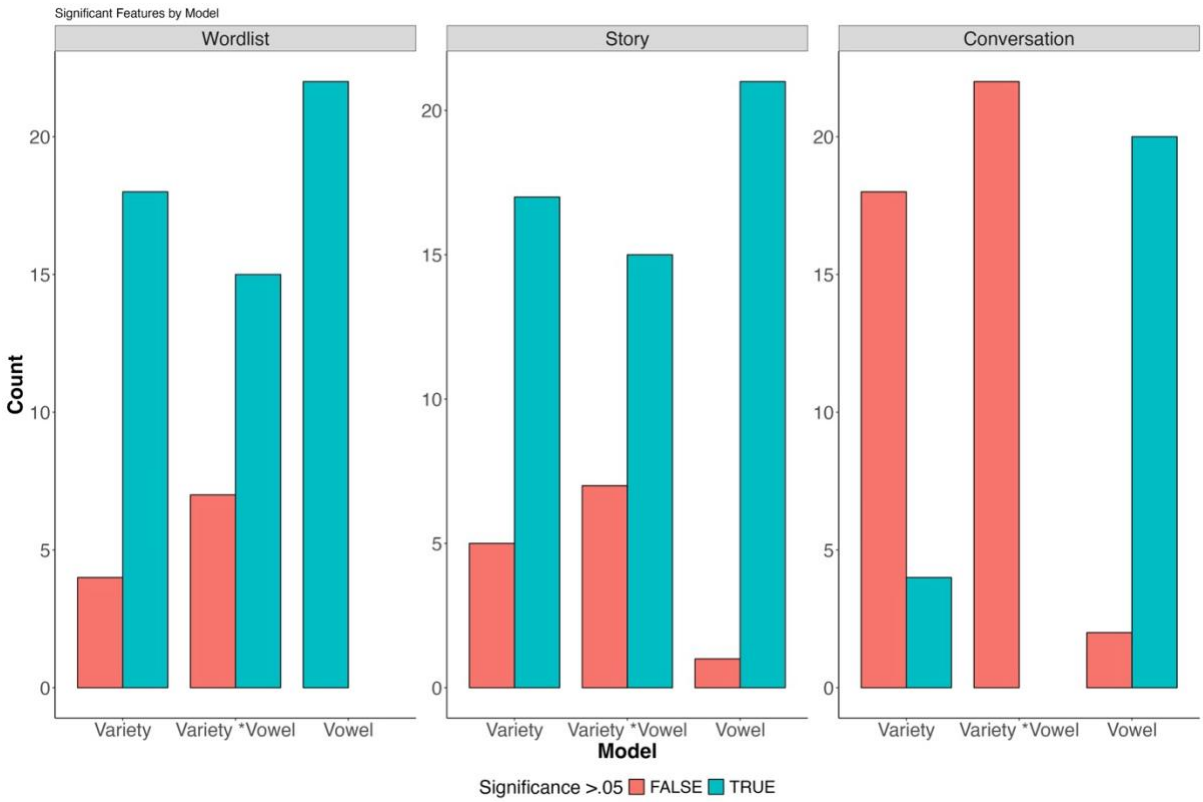


Figure 7.2 Significance levels of different models. The figure depicts a bar chart in which the significant count of different models is expressed as true or false. The chart's x-axis depicts the models as well as the number of true or false results.

7.8.2 Impact of including spectral measures with formant centre frequencies for speaker discrimination

The next step was to analyse these measures to assess their role in spectral measure analysis for individual speakers. This analysis was conducted with the help of linear discriminant analysis (LDA). LDA determines the prediction rate of an individual measure. It looks for the underlying vectors that are most useful for classifying speakers (Fisher, 1938; Martinez & Kak, 2001) based on two main assumptions: data is normally distributed, and each class has equal covariance matrices. The data was z-transformed for normalisation, and 'Levene's', 'Bartlett's' and 'BoxM' tests were conducted to test the homogeneity of the data. Both correlations and covariance of the variables were tested. To verify the correlation between these measures, a correlational analysis was conducted with the help of 'ggcorrplot' (Kassambara & Patil, 2023) for R. This analysis also included the centre formant frequencies of each formant. Figure 7.3 shows the correlations between the spectral measures. Any outcome below 0.6 was considered collinear and discarded from being in the same model. Formant frequencies are significantly correlated with their corresponding spectral peaks, thus suggesting that they will cause multi-collinearity issues if included in the same model (Boedeker & Kearns, 2019). Formant amplitude values were also collinear with each other. This step helped determine which individual measures can be combined in the same model.

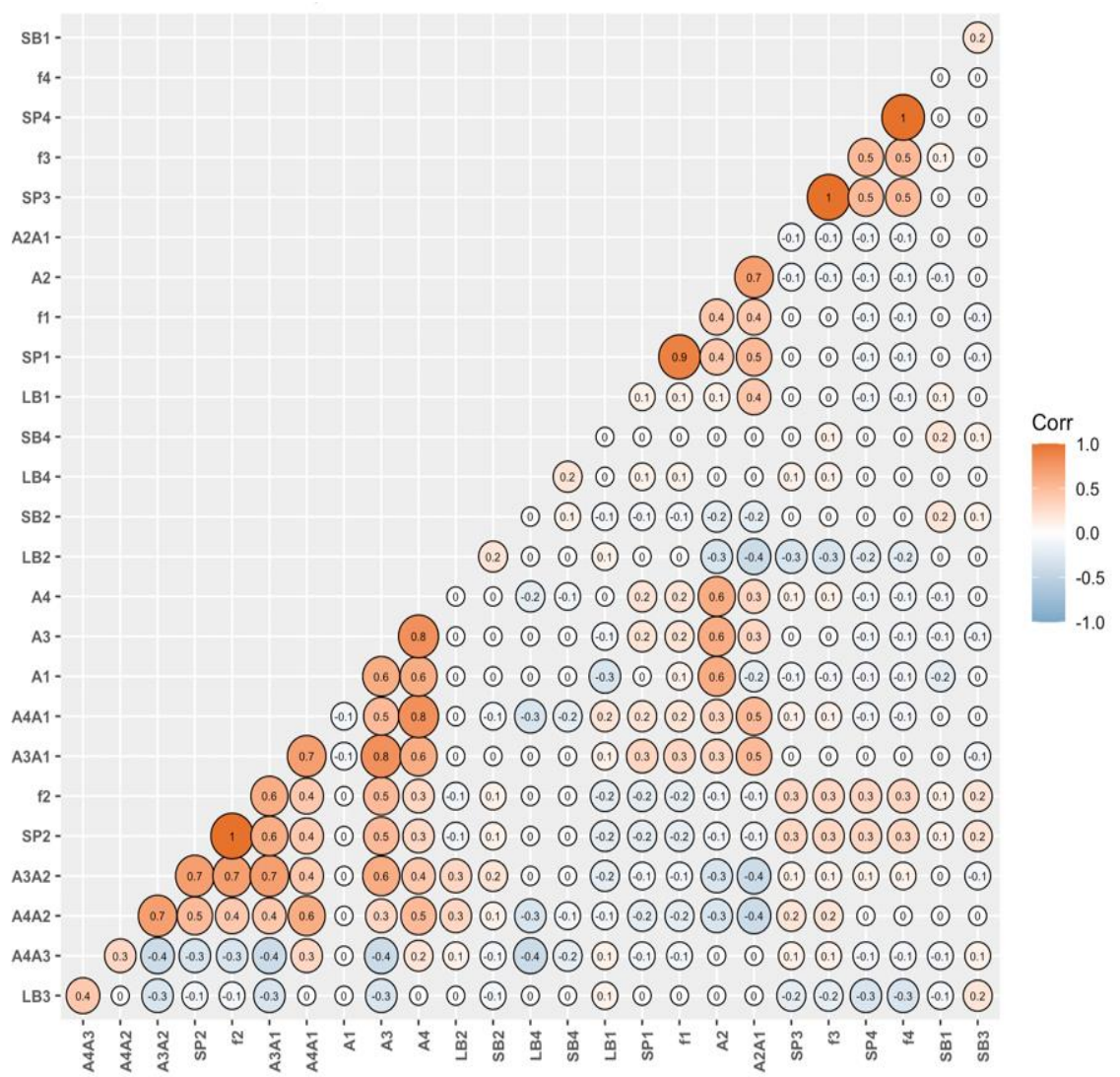


Figure 7.3 A correlation plot, visualizing the relationships between 22 individual spectral measures and four centre formant values (F1, F2, F3 and F4). The correlation is tested for each feature against the other, where the colour of the circle represents the range of correlation, while its size represents the degree of correlation, i.e., the larger the circle, the more correlated/uncorrelated the value (Kassambara & Patil, 2023).

Once the LDA assumptions were satisfied by inputting the extracted data for all three modes of data collection separately, the analysis was run in R using the 'mass' package (Ripley et al., 2023). The analysis looked at three questions: assessing if including spectral measures with formant centre frequencies improved the model (RQ2), assessing the speaker-discriminatory power of individual spectral measures, and assessing the speaker-discriminatory power of multiple combinations of spectral measures (RQ2.1).

As mentioned in section 7.4.1, formant centre frequencies alone have shown to be a significant measure for speaker discrimination. The next step was to verify if including the measures selected for study also impacted the classification rates for the model. Table 7.9 shows the classification rates times above chance level for three different varieties. The table also shows that these classification rates (CRs) for three different modes of data elicitation, wordlist , story and conversation. The first row demonstrates the times above chance level for centre formant frequencies alone, and the impact of each measure is tested later by adding them with formant values. The table shows that by adding just one feature at a time the CRs increased up to ten times above the chance level.

Table 7.1 Times above chance of classification rates extracted from LDA shown for centre formant frequencies alone and with spectral measurements added one at a time.

	Wordlist	Story	Conversation
F1+F2+F3+F4	8.1	6.5	4.6
F1+F2+F3+F4+A1	9.1	8.4	5.2
F1+F2+F3+F4+A2	9.7	8.0	5.2
F1+F2+F3+F4+A3	9.5	8.0	5.5
F1+F2+F3+F4+A4	10.3	8.6	5.1
F1+F2+F3+F4+SB1	8.9	7.1	5.2
F1+F2+F3+F4+SB2	8.8	6.6	5.2
F1+F2+F3+F4+SB3	8.6	6.9	5.0
F1+F2+F3+F4+SB4	8.5	7.5	4.7
F1+F2+F3+F4+SP1	8.4	6.6	4.9
F1+F2+F3+F4+SP2	8.1	6.9	5.0
F1+F2+F3+F4+SP3	8.4	6.6	5.3
F1+F2+F3+F4+SP4	8.3	6.7	5.1
F1+F2+F3+F4+LB1	8.7	6.9	5.0
F1+F2+F3+F4+LB2	8.4	6.5	4.9
F1+F2+F3+F4+LB3	8.7	6.9	5.0
F1+F2+F3+F4+LB4	8.5	6.7	4.9
F1+F2+F3+F4+A2A1	8.6	6.2	5.0
F1+F2+F3+F4+A3A1	9.2	6.8	5.1
F1+F2+F3+F4+A4A1	10.6	7.1	5.1
F1+F2+F3+F4+A4A2	9.2	7.1	4.9
F1+F2+F3+F4+A4A3	9.7	6.9	5.1
F1+F2+F3+F4+A3A2	8.3	6.6	4.9

7.8.3 Speaker discriminatory power of an individual spectral measures

Table 7.2 shows each spectral measure's 'times greater than the chance factor'. The chance level (CL) for each category changed according to the number of tokens per participant. Overall, the CRs for the conversational data were higher than CRs for the rest. The chance level of CRs for individual measures drastically increased the predictability of the model. The measures discriminated the speakers for wordlist data for a minimum of 2.8 times above the chance level. This increased for the story and conversation data. Although no measure performed well for all three modes of data elicitation, the results suggest that formant amplitude and spectral peaks were better discriminators than the rest of the measures. A summary of the results is presented in Table 7.2. The table shows the values of CR times above the chance threshold. For example, for wordlist data, the highest performing measure was spectral peak derived from F3, which produced the CR 5.8 times higher than chance (which was 2.22).

Table 7.2 Summary of individual measure performance for LDA

Data Type	N. of Participant	Chance level	Mean	Best measure	Worst measure
Wordlist	45	2.2	4.3	SP3(5.8)	A3A2 (2.8)
Story	44	2.2	4.1	A4 (5.1)	SB3 (3.2)
Conversation	25	4.0	5.5	SP4 (7.2)	A3A1 (4.6)

Table 7.3 lists the eight most effective measures for each mode of data collection. This data did not show a clear trend, however several measures performed well for each mode. For example, amplitudes and spectral peak values outperformed the other characteristics.

Table 7.3 Eight best-performing measures for every mode of data collection

Wordlist	Story	Conversation
SP3	A4	SP4
A2	A2	SP3
SP4	A1	A1
A4	SP3	SP1
A1	LB1	A2
SB1	A4A3	A2A1
A3	SB1	LB3
A4A1	A4A1	LB1

7.8.4 Speaker discriminatory power of the combinations of spectral measures

Table 7.4 Summary of combinations of spectral measure performance for LDA (same measures)

Model No.	Measures	Times above chance		
		Wordlist	Story	Conversation
1.	A1+A2+A3+A4	8.0	6.7	6.2
2.	SP1+SP2+SP3+SP4	8.0	6.3	10.0
3.	SB1+SB2+SB3+SB4	5.0	4.6	4.5
4.	LB1+LB2+LB3+LB4	6.3	5.4	6.5
5.	A2A1+A3A1+A4A1+A3A2+A4A2+A4A3	5.3	6.1	5.9

Table 7.5 Summary of combinations of spectral measure performance for LDA (best measures where collinearity was ignored)

Model No.	Measures	Times above chance		
		Wordlist	Story	Conversation
6.	SP3+A2+SP4+A4+A1+SB1 +A3+A4A1	11.56	6.48	5.52
7.	A4+A2+A1+SP3+LB1+A4A3+SB1+A4A1	10.54	6.48	4.92
8.	SP4+SP1+SP3+A1+A2+LB3+LB1+A2A1	10.54	7.02	5.88

Table 7.6 Summary of combinations of spectral measure performance for LDA (best measures where collinearity was accounted for)

Model No.	Measures	Times above chance		
		Wordlist	Story	Conversation
9.	SP3+A2+SP4+SB1+A4A1+LB2+LB2+SP1	10.20	6.48	5.88
10.	A4+SP3+LB1+A4A3+SB1+A2A1+LB4+LB3	9.86	5.40	5.04
11.	SP4+A1+SP1+SP3+A2+LB3+SB3+LB1	10.88	7.56	6.00

The next step was to assess the performance of the combinations of the measures with the help of LDA. An analysis was conducted with combinations of the same measures for different formants, followed by a combination of relative amplitudes. Table 7.4 provides the CRs for these measure combinations and relative amplitudes. SPs performed better than the other measures in this step (the issue of collinearity was ignored at this step as the sample size was substantial). Relative amplitude (amplitude differences) values performed as well as the other measures. LB performed better than SB.

The next step was to create combinations of the best-performing measures. For this step, two different methods were employed: Best-performing variables where collinearity was ignored because, although the collinearity was detected for amplitudes, the variation of inflation tested for the model was < 5 (Weisberg, 2005). The second method was to avoid collinear variables altogether, by removing them from the model.

Table 7.5 and 7.6 show the two different kinds of analysis. Although removing the collinear measures decreased the outcome of CRs for some models (e.g., 9_wordlist 10_story), it also positively impacted other models (e.g., 11).

7.8.5 Speaker discriminatory power of the combinations of best-performing spectral measure and formant

The third step of discriminatory power analysis of the spectral measures was to test these measures with centre formant frequencies. Including centre-formant frequencies increased the CRs above chance level up to twelve times (with amplitude). Table 7.7 also suggested that with formant frequencies included, wordlist data performed better than the other two modes. However, this performance was not significant and un-generalisable because the sample size had drastically reduced for the conversation data.

Table 7.7: Combinations of spectral measures and formants

Model No.	Measures	Times above chance		
		Wordlist	Story	Conversation
12.	A1+A2+A3+A4+F1+F2+F3 +F4	12.24	7.83	6.36
13.	SB1+SB2+SB3+SB4+F1+F2+F3+F4	7.48	5.13	5.28
14.	LB1+LB2+LB3+LB4+F1+F2+F3+F4	8.50	7.02	5.52
15.	SP3+A2+SP4+SB1+A2A1+F1+F2+F3+F4	10.20	5.94	6.00
16.	A4+SP3+LB1+A4A3+F1+F2+F3+F4	11.22	6.75	6.60
17.	SP4+A1+SP1+SP3+F1+F2+F3+F4	9.52	5.67	6.36

7.8.6 The discriminatory power of spectral measures for different vowels

The next step was to assess the role of individual vowels and their impacts on the LDA CRs of the measures. Participant numbers were further reduced to accommodate the prerequisite n-1 for LDA. For instance, the tokens for vowel /ʊ/ reduced the LDA vectors to five due to the lack of utterances of this vowel in conversation and story data. Filtering vowels for the model presented a clear pattern. Some vowels helped with the CRs better than others. Vowel /ɪ/, /a:/ and /e/ consistently produced higher classification rates for every measure than the others. Whereas vowels /o/ and /i:/ provided consistently lower CRs. Overall, high vowels had lower CR rates than the low vowels (with one exception of vowel /ɪ/). Figure 7.4 shows the results of the spectral measure analysis for individual vowels. Spectral measures extracted from the higher formants such as SP4, SP3 or A3, A4 performed better for vowel-dependent models than the rest. However, there was no clear pattern observed.

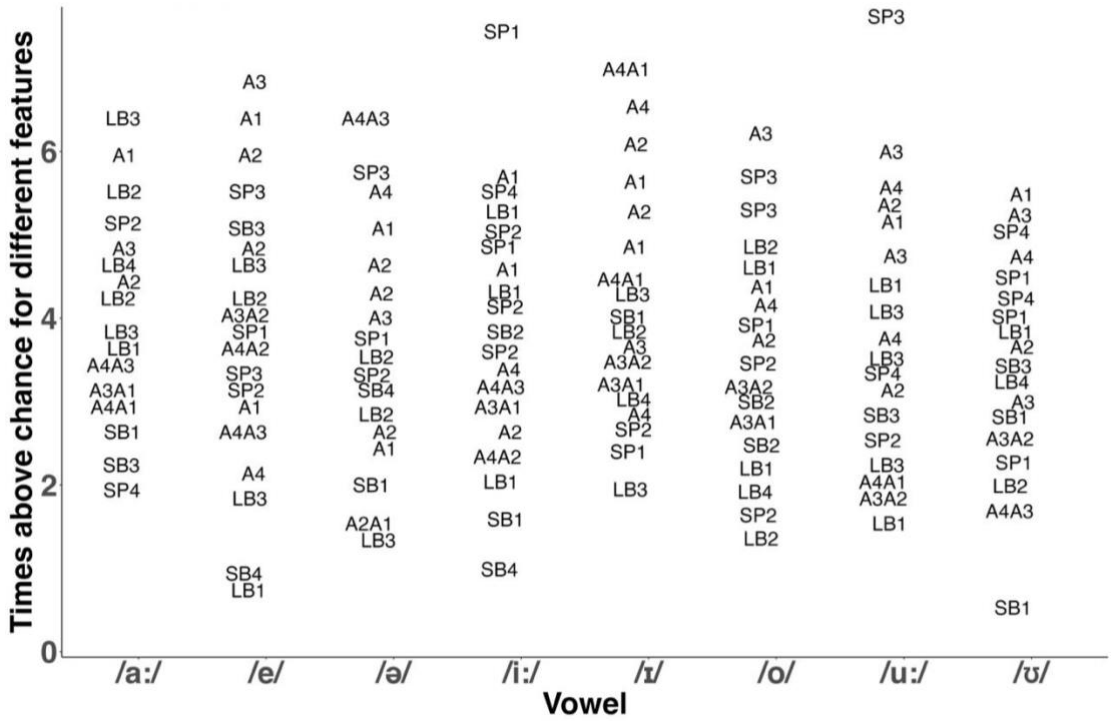


Figure 7.4 Spectral measure analysis for individual vowels.

Once the relevance of the individual measures and the impact of vowels on these measures had been determined, the next step was to test the same for best-performing measure combinations. These combinations were based on the eight best-performing measures for each data elicitation mode. The hypothesis was that by combining eight best-performing measures would considerably increase the CRs.

SP3+A2+SP4+SB1+A4A1+LB2+LB2+SP1 (C1)

A4+SP3+LB1+A4A3+SB1+A2A1+ LB4+LB3 (C2)

SP4+A1+SP1+SP3+A2+LB3+SB3+LB1 (C3)

Because of the reduced sample size for this step, only the measures determined to be non-collinear were used in the models. Three models were created based on the best-performing measures for each data type. Wordlist data was the best-performing mode of data elicitation for this step. The CRs increased up to 10.5 times for the vowel /e/. The same vowel performed best for all three models for wordlist data. For the following two modes of data elicitation, vowel /i:/ provided the highest classification rates. These results were different from individual vowel performances where vowel /i:/ provided the lowest classification rates. Back vowels and

/ə/ consistently provided lower CRs (Table 7.8). The mean classification rates were also two times higher for wordlist data.

Although the CRs for story and conversation were lower than that of wordlist, it is noteworthy that the numbers represented here are the times above the chance level, thus illustrating how well every model performs above chance. The values for story and conversations increased a minimum of two times above the chance level, suggesting the significance of these models for classification rates.

Table 7.8 Performance of spectral measures of vowels times above the chance level

Vowel	Measures	Times above chance		
	Models	Wordlist	Story	Conversation
/aː/	C1	6.7	3.1	3.8
	C2	6.3	3.8	3.5
	C3	6.7	3.1	3.3
/e/	C1	10.5	5.0	4.1
	C2	8.7	6.0	4.3
	C3	10.0	5.6	4.0
/ə/	C1	9.0	4.2	1.5
	C2	6.9	3.3	1.2
	C3	8.4	4.3	1.3
/o/	C1	4.8	5.8	5.4
	C2	4.6	5.2	5.6
	C3	4.6	4.4	6.0
/uː/	C1	7.0	2.3	1.2
	C2	6.6	2.0	1.4
	C3	7.2	2.0	1.3
/ʊ/	C1	3.3	2.0	2.0
	C2	3.6	2.0	2.0
	C3	3.9	1.5	2.0
/iː/	C1	5.4	6.2	7.2
	C2	3.8	6.0	6.1
	C3	4.7	6.5	7.7
/ɪ/	C1	9.9	4.8	4.2
	C2	7.5	4.6	2.4
	C3	8.4	5.8	3.8
	Mean	6.6	4.1	3.5

7.8.7 The discriminatory power of spectral measures for different varieties

Table 7.9 Classification rates of centre formant frequencies and spectral measures together and separated for different varieties.

	Brahmin			Jaat			Bishnoi		
	Wordlist	Story	Conversation	Wordlist	Story	Conversation	Wordlist	Story	Conversation
F1+F2+F3+F4	4.9	4.9	2.9	3.2	3.2	2.4	3.9	3.3	1.5
F1+F2+F3+F4+A1	6.5	5.1	2.2	4.0	2.8	5.2	5.7	3.1	6.8
F1+F2+F3+F4+A2	6.8	4.9	2.1	3.1	4.1	2.5	4.8	5.0	2.7
F1+F2+F3+F4+A3	6.5	4.7	2.4	3.0	3.6	2.5	4.4	5.1	2.8
F1+F2+F3+F4+A4	7.1	5.0	2.3	4.0	4.0	2.6	5.1	5.1	2.5
F1+F2+F3+F4+SB1	6.3	4.4	2.3	3.4	3.9	2.4	4.9	4.2	1.8
F1+F2+F3+F4+SB2	5.9	4.0	2.3	3.0	3.0	3.5	2.5	4.0	3.5
F1+F2+F3+F4+SB3	5.9	4.0	2.1	3.2	3.7	2.4	4.1	3.2	3.7
F1+F2+F3+F4+SB4	6.0	3.9	2.0	3.2	4.2	2.5	4.2	3.4	1.9
F1+F2+F3+F4+SP1	5.9	4.2	2.1	3.0	3.6	2.5	4.2	3.1	1.9
F1+F2+F3+F4+SP2	5.9	3.7	1.9	3.2	3.5	2.5	4.3	3.4	1.8
F1+F2+F3+F4+SP3	6.0	4.1	2.1	3.1	3.5	2.5	4.3	3.3	2.3
F1+F2+F3+F4+SP4	5.9	4.3	2.0	3.0	3.5	2.5	4.0	3.4	1.8
F1+F2+F3+F4+LB1	6.4	3.6	2.2	3.7	3.7	2.6	4.3	3.4	1.8
F1+F2+F3+F4+LB2	6.2	3.9	2.1	3.1	3.4	2.6	4.1	3.5	1.8
F1+F2+F3+F4+LB3	6.1	4.1	2.2	3.1	3.3	2.5	4.2	3.9	1.9
F1+F2+F3+F4+LB4	6.0	4.2	2.2	3.4	4.0	2.5	3.9	3.4	2.0
F1+F2+F3+F4+A2A1	4.3	3.5	3.4	4.1	3.3	3.2	4.1	3.0	3.3
F1+F2+F3+F4+A3A1	4.1	3.1	3.2	4.0	2.9	3.1	4.1	3.9	2.9
F1+F2+F3+F4+A4A1	4.1	3.0	3.1	4.0	3.1	3.0	4.0	3.3	3.2
F1+F2+F3+F4+A4A2	4.3	3.3	3.6	4.0	3.1	3.1	4.0	3.0	3.0
F1+F2+F3+F4+A4A3	3.9	3.3	3.2	3.9	3.0	3.1	3.9	2.9	3.1
F1+F2+F3+F4+A3A2	3.9	2.9	3.1	4.0	3.3	3.2	4.1	3.0	3.1

The impact of spectral measures on different varieties was tested with the help of two separate models. The first model (M1) had spectral measure CRs for all varieties together, i.e., the entire data was treated as a single language and variety distinctions were removed. The second model (M2) used variety subsets for vowel averages.

Figure 7.5 shows that the average CRs of vowels decreased drastically once the variety subsets i.e. M2 was created, which could be explained by the inevitable reduction in the number of participants. The next step was to assess the individual performance of spectral measures for each variety.

This step was divided into two parts. The first part evaluated the performance of individual spectral measures for each variety and the second part looked at the models (C1-C3) created in section 7.8.6.

Figure 7.6 shows the performance of every spectral measure for each variety. Formant amplitudes showed the highest performance for TAC, followed by bandwidth measurements. Overall, the averages for all three varieties were similar for all three data types. Table 7.10 depicts the averages acquired for each type of data and the highest-performing measure for each vowel of each data type of the three varieties. Again, no clear pattern was observed for any vowel performing best or worst. However, for every vowel subset, the CR increased to a level of at least 2.5 times above chance for every variety. This suggests that creating subsets for each variety and type bears significant advantages for classifying speakers. Front vowels showed slightly higher CRs than back vowels, but further testing is needed.

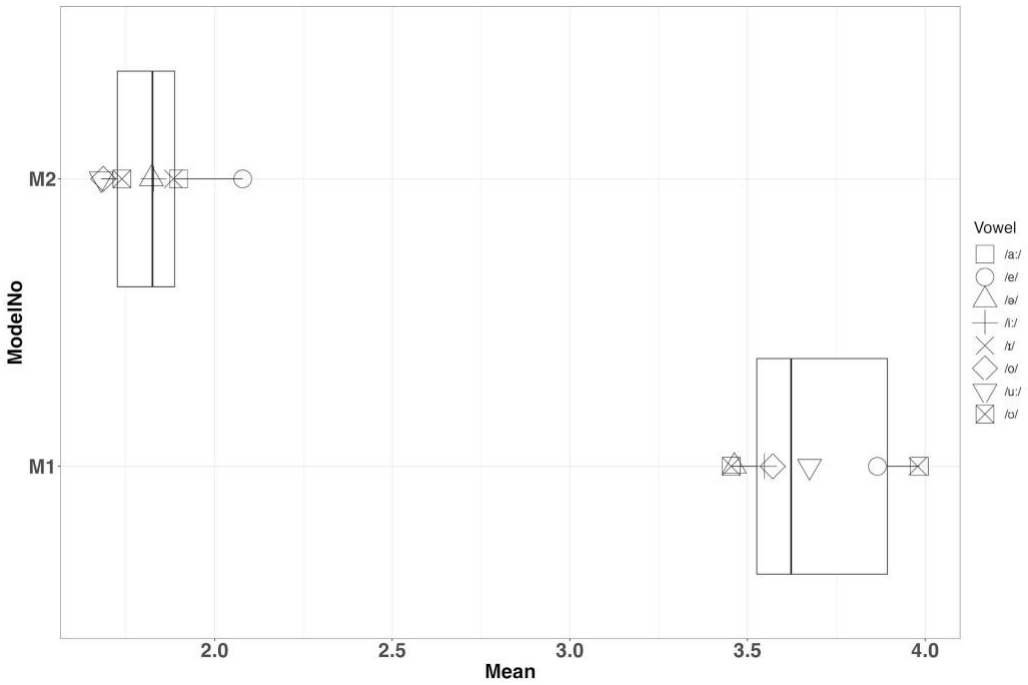


Figure 7.5 Difference between CRs for M1 and M2.

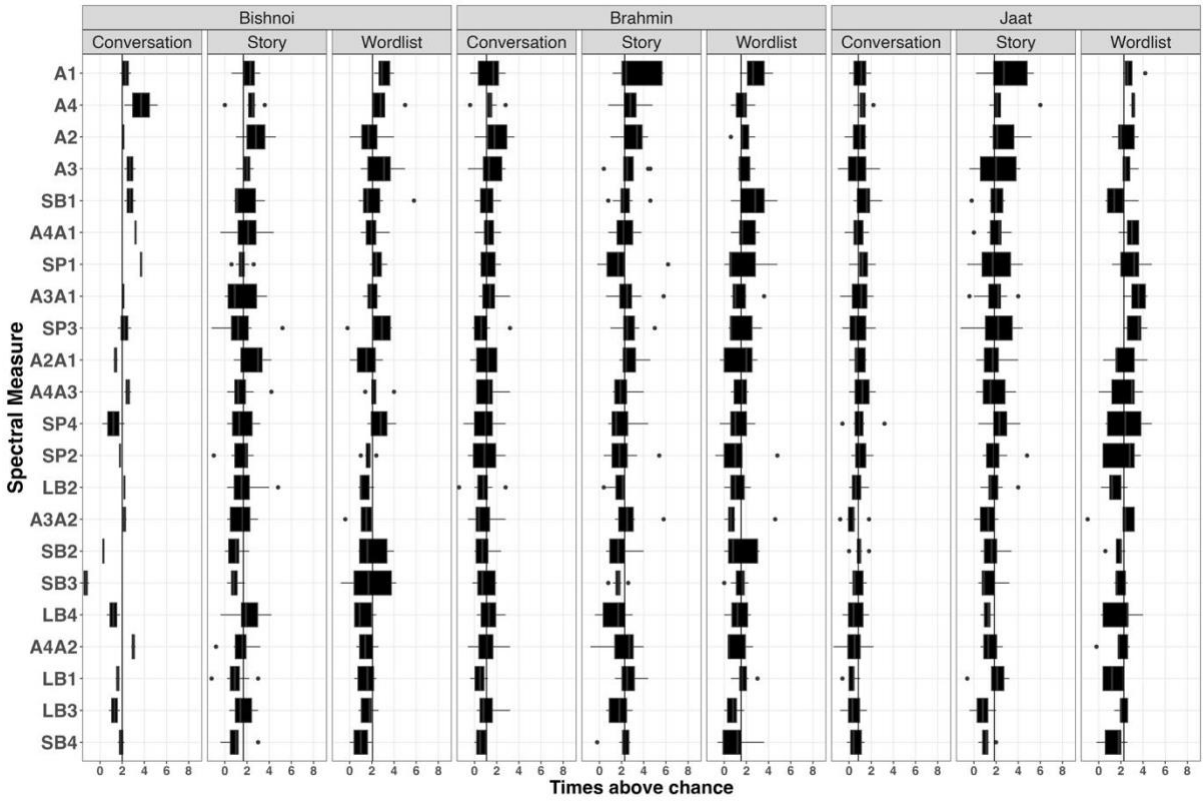


Figure 7.6 Individual performance of spectral measures for each variety (top-labels) and each type (bottom-labels).

Table 7.10 Average performance of varieties for each type for every vowel

Variety				**Brahmin**				**Jaat**				**Bishnoi**			
Type	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean
Mean	1.8	2.1	1.5		2.1	1.9	1.4		2.0	1.9	1.9	
/a:/	2.5	2.5	2.4	2.5	3.4	3.4	2.4	3.1	2.7	2.6		2.6
/e/	3.4	2.9	2.8	3.0	3.1	3.1	2.1	2.8	7.6	3.6		5.6
/ə/	2.0	4.1	2.6	3.0		4.0	1.8	2.9	3.5	2.5		2.9
/i:/	3.2	3.8	2.2	3.1		2.4	2.6	2.5		2.7	2.9	2.8
/ɪ/	3.4	3.3	1.8	2.9	2.8	3.37	2.2	2.8	2.9	2.9		2.9
/o/	2.7	3.7	2.2	2.9		2.7	1.6	2.2		2.4	3.6	3.0
/u:/	3.2	2.5	1.7	2.4	2.5	2.5	2.2	2.4	3.9	3.4		3.6
/ʊ/		3.0	2.0	2.5	3.1	3.4	1.7	2.7	2.9	2.2		2.6

The table 7.10 shows the best CR rates above chance level for each vowel by variety and elicitation mode. CR rates ranged from 1.4 to 7 times above chance. Rates could not be calculated for some vowels due to insufficient tokens.

For Brahmins, wordlist had the highest average CR at 4.1 times above chance. Jaats showed equal increases of 3.4 times chance for both wordlist and conversation. Bishnois exhibited the

top rate during conversation at 7.6 times chance. However, this estimate is less reliable as only 2 vowels had sufficient tokens for analysis in this elicitation. Overall, performance patterns varied by vowel and variety across modes. Wordlist tended to enable the best discrimination for Brahmins and Jaats. But higher CR rates emerged for Bishnois during conversation, albeit with limited vowel data. Further analyses of the relationship between spectral measures and CR rates by mode would clarify optimal elicitation approaches for speaker discrimination within each variety.

The next step was to analyse variety-specific differences for individual models (see section 7.8.6). C1 outperformed the other two models drastically, suggesting the best possible combination of spectral measures is: SP3+A2+SP4+SB1+A4A1+LB2+LB2+SP1.

Table 7.11 Variety-specific differences for individual models

Variety				**Brahmin**				**Jaat**				**Bishnoi**
Type	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean
C1	2.7	3.9	3.5	3.4	3.5	1.9	3.3	2.9	2.2	2.7	4.6	3.1
C2	1.8	3.8	3.2	2.9	3.5	2.0	3.6	3.0	1.7	2.3	3.8	2.6
C3	1.8	3.8	3.2	2.9	3.5	2.0	3.6	3.0	1.7	2.3	3.8	2.6
Mean	2.1	3.8	3.3		3.5	1.9	3.5		1.8	2.4	4.1	

7.9 Summary

We began with the following research questions, which for ease of reference, are reproduced here:

RQ1) Are spectral measure values impacted by variety, vowel and mode of data elicitation?

A1 Yes, but the Linear mixed model ANOVA conducted to test the roles of vowel, variety and mode of data elicitation showed that only vowels and mode of data elicitation significantly affected the values. The results for variety depended on the tested measure.

RQ2 Can including spectral measures with formant centre frequencies help distinguish between individual speakers in an acoustic analysis?

A2 Yes, the overall CRs surpassed the chance level by 4.9 times, suggesting that including spectral measures with formant centre frequencies does help with distinguishing speakers.

RQ2.1) Which spectral measures and combinations of spectral measures are most effective?

A2.1 Individual measures: The overall results differed for each formant in respect of individual spectral measures. Some measures provided CRs as high as seven times above the chance level, all measures yielded CRs at least 2.8 times above the chance level. Amplitude and spectral peak provided better CRs than either the LPC-derived bandwidth or spectrally derived bandwidth. LPC bandwidth performed better than spectral bandwidth.

Combination of measures: The combination of spectral measures significantly improved the models' performance. The CRs increased from 2-6 times to 5-9 times for four measure combinations and 12 times for combinations with eight spectral measures.

RQ3 Are there any factors that impede or facilitate spectral measures' discriminant values? If yes, what are these factors?

A3 Yes, the initial ANOVA carried out to test the effect of variety, vowel and mode of data elicitation showed that all these factors had an impact on the discriminant values of the spectral measures.

So, after further dividing these factors we could answer the following questions:

RQ3.1 Which vowels or subsets of vowels yield the highest CR results when spectral measures are applied?

A3.1 Vowels /u:/, /e/, /a:/ and /ɪ/ provided higher CRs than vowels /i:/, /o/, /ə/ and /ʊ/ (Figure 7.4). The performance of the front vowels was slightly better than the back vowels. For the combination models, the best-performing model was C1.

RQ3.2 Which speech styles provide the highest CRs when spectral measures are applied to them?

A3.2 The average performance of each model was 4.9 (wordlist), 4.4 (story) and 4.8 (conversation), thus suggesting that including up to eight of the best-performing spectral measures in a model increases the CRs at least four times above the chance level, regardless of the data elicitation mode.

RQ3.3 For which varieties do spectral measure analysis work best?

A3.3 The models performed similarly for all three varieties, showing that no spectral measure performed better on any particular variety; they were roughly equivalent.

7.10 Discussion

We now turn to some possible explanations of our findings:

7.10.1 Individual spectral measures

7.10.1.1 Bandwidth

The performance of bandwidth, though significant, was less so than the other two variables (amplitude and spectral peak). This could be because bandwidth may be related to vocal tract losses rather than vocal tract shapes (Fant, 1972; Millhouse et al., 2002). These losses can be treated as frequency or area functions of the vocal tract. But vocal tract loss identified by bandwidth analysis was proven to be less efficient as a stand-alone measure. The explanation for these results could be that the recording quality was significantly impacted by the non-laboratory settings of participants' houses, thus failing to capture the minute vocal tract losses such as radiation or heat conduction. Their failure could have impacted the extraction process for bandwidth. Since forensic voice analysis often relies on non-pristine recordings, the inability to capture subtle formant patterns such as bandwidths in this study's non-laboratory setting indicates practical case conditions may impede accessing the fine-grained spectral details theoretically useful for speaker discrimination. Real-world ambient noise likely degrades sensitivity to minute vocal tract resonances informative for identification. This necessitates developing parameters robust to spectral detail loss and cautions forensic examiners against over-interpreting bandwidth findings that environmental factors may obscure. Additional research on distinguishing speakers from acoustically-degraded evidence is needed to enhance forensic technique validity when applied to uncontrolled case recordings that filter out microscopic yet potentially relevant acoustic patterns. Another explanation could be that extracting bandwidth, as suggested in section 7.4.3, is a highly complex process, and often the measurements are only a very close approximation of the actual value.

7.10.1.2 Spectral Peak

The most significant measure was the spectral peaks extracted from the closest maximums from the formant frequencies. This was expected as, most of the time, the highest maximum from the formant would be aligned with the formant itself, i.e., closely aligned with the formant centre value.

7.10.1.3 Amplitude

Formant amplitude was the second-best performing variable after spectral peaks. As Fant et al. (1963) suggested, amplitude one of the measures that can signify vocal tract shape, providing the initial hypothesis that amplitude would be a significant measure for speaker discrimination. The individual spectral measure results supported this hypothesis. For each mode of data collection, amplitude alone was proven to increase the CRs drastically. Although the results suggested a slightly better performance from higher formant amplitudes, the overall difference between high and low formant amplitude was not significantly different. This result could be an essential finding; as discussed earlier, it is not always possible in forensic casework for a recording to contain data from higher frequency regions because of the bandwidth limitation in telephonic conversations.

7.10.2 Combinations of spectral measures

The combinations that performed best consisted of amplitude and spectral peak measurements. This is in line with the literature, as spectral peaks are essentially formant centre frequencies in most cases. Moreover, amplitude values have a direct correlation with formants as they were extracted as a peak amplitude based on the sound pressure level of the nearest formant peak, thus making the value formant peak-dependent. The combinations that accounted for this collinearity also provided high CRs, suggesting that adding these variables together might benefit the model for manual speaker discrimination process.

7.10.3 Which vowels?

Although front vowels performed slightly better than the back vowels, there was no clear pattern that emerged from the vowel analysis. There are two possible explanations for this; the data was collected in non-ideal situations, thus many minute characteristics of the measures

analysed here could have been lost. This could also have affected the close-grained representations of vowel formants, especially for story and conversational data. The second reason could be that creating subsets of vowels reduced the number of variables tested for the study drastically, thus not providing enough data to create a distinct pattern.

7.10.4 Which mode of data elicitation?

Out of three data elicitation techniques selected here, wordlist performed the best, which is not surprising as this method provided a steady state of vowel extraction. There is more control over the phonological environment of consonants preceding and following the vowels. Conversational data or spontaneous data is often associated with increased vowel reduction, thus directly impacting harmonics and resonance frequencies (Aylett & Turk, 2006). Speech signal is also impacted by the amount of attention paid to the speech, less vowel reduction and centralisation happening towards the more attentive speech (in this case wordlist) and vice versa (Picheny et al., 1986).

7.10.5 Which variety?

Individually, amplitude performed better than the other measures, this could again be linked to the high collinearity between amplitude and formant centre frequencies. However, the measures for combinations and vowels did not allow for a clear distinction between the varieties, suggesting that these measure despite being able to improve the model by four times, they are not particularly variety sensitive. Forensically, this can potentially be interpreted as the measures carrying more inter-speaker than inter-variety variations. Hence, this could be useful for speaker discrimination work.

7.11 Limitations

As mentioned earlier in Section 7.4, bandwidth and amplitude are highly affected by the quality of data collected, and since the data was collected in speakers' houses, the impact of various surrounding parameters should be addressed i.e., controlled for in any further study. In other words, recording environment should be the same for all participants. The second limitation is that because telephone transmitted speech typically has an upper bandwidth frequency of 4 KHz, it would not be possible to apply spectral measure analysis to (the absent) F4. The data collection also had some significant limitations, as the number of tokens drastically reduced

for conversation speech because of the *n-1* limitation associated with LDA. This could be eliminated by collecting and analyzing data from a larger sample. But at the same time, extracting these measures was time-consuming and labour-intensive, requiring a more automated but dependable system. The study was also limited to female speakers from a small region; testing needs to be conducted on male and age-dependent speech to have a more robust and replicable model. The data also needs to be tested on different languages, especially those from different language families, to see of the results can be replicated.

7.12 Implications for Research

The results for 45 speakers and their spectral measure analysis indicate that including amplitude, bandwidth and spectral peaks in forensic casework might significantly increase the correct CR for an individual speaker. Although further testing is needed to assess the applicability of this method to different languages, channels and a bigger database, it is suggested that including these measures along with already established auditory-acoustic parameters could make speaker classification or discrimination casework more accurate.

8 Chapter 8 Presented as Article 3

8.1 Research Degree Thesis Statement of Authorship

University of York

York Graduate Research School

Candidate name	Nikita Suthar
Department	Department of Language and Linguistic Science
Thesis title	Within-formant spectral feature analysis for forensic speaker discrimination casework: A study of 45 Marwari monolinguals from Bikaner, India

Title of the work (paper/chapter)	Enhancing forensic speaker discrimination: a comprehensive spectral feature analysis of Marwari vowels using within-formant measures and spectral moments.	
Publication status	**Published**	
	Accepted for publication	
	Submitted for publication	
	Unpublished and unsubmitted	*
Citation details (if applicable)	NA	

Description of the candidate's contribution to the work	Conceptualisation, literature review, data collection and analysis, writing and manuscript preparation, citation, and references
Percentage contribution of the	90%

candidate to the work	
Signature of the candidate	Nikita Suthar
Date (DD/MM/YY)	25th September 2023

Co-author contributions*

By signing this Statement of Authorship, each co-author agrees that:

(i) the candidate has accurately represented their contribution to the work;

(ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co-author	Prof. Peter French
Contact details of co-author	peter.french@york.ac.uk
Description of the co-author's contribution to the work**	Part of conceptualising, contribution to interpretation of findings, editing
Percentage contribution of the co-author to the work	10%
Signature of the co-author	
Date (DD/MM/YY)	27th Sept 2023

8.2 Title: Enhancing Forensic Speaker Discrimination: A Comprehensive Spectral Feature Analysis of Marwari Vowels Using Within-Formant Measures and Spectral Moments

8.3 Abstract

Centre formant frequencies, trajectories, LTFDs, and, to a lesser extent, bandwidths have been the primary focus of formant-based forensic speaker classification and discrimination investigations. The model used in the current article incorporates a linear discriminant analysis (LDA)-based model that makes use of within-formant characteristics. Various spectrum measurements such as formant bandwidths (estimated using spectral and linear predictive coding (LPC) approaches), amplitude, spectral peaks, and the first four primary spectral moments (centre of gravity, standard deviation, skewness, and kurtosis) are among these properties. By including these factors, correct classification rates may be elevated noticeably, up to 10 times higher than with previous approaches. The research is based on recordings of 45 female Marwari monolingual speakers, 15 from each of three caste dialects - Bishnoi, Jaat, and Brahmin All speakers were from the Rajasthan region of India's Bikaner district. Wordlist, story, and conversation were the three methods used to gather the data.

Keywords*: within-formant features, spectral moments, amplitude, bandwidth, spectral peaks.*

8.4 Introduction

Formant analysis has played a significant role in speaker identification work for many decades (e.g., Cao & Dellwo, 2019; Fant, 1971; McDougall & Nolan, 2007; Stevens, 2000). Formants represent the prominent resonance frequencies. Their centre frequencies are constrained by the phonology of the language being spoken i.e., meeting articulatory 'targets' for the vowels within the system, However, the centre frequencies, and other more 'subtle' frequencies are also shaped by individual biology. They reflect the geometry and configuration of an individual speaker's vocal tract. This latter characteristic of formants makes them useful for forensic speaker comparison. (see e.g., Ingram et al., 1996; Jessen, 1997, 2008; Liepins et al., 2020; Nolan, 1983).

The role of formant-based analysis in speaker comparison casework is discussed in *inter alia* (e.g., Cao & Dellwo, 2019; Cavalcanti et al., 2021; Gold & French, 2019; McDougall & Nolan, 2007; Nolan, 1983; Yang et al., 1996).

Researchers have also investigated sub-vocalic segments, such as spectral moments of '*isochuncks*' extracted from instances of the same speech produced by a speaker at different times (Eriksson, Cepeda, Rodman, McAllister, et al., 2004; Yang et al., 1996), within-formant spectral moments (Suthar & French, 2023a), amplitudes (Mitra et al., 2012), and bandwidths (Gonzalez-Rodriguez, 2011). The approach of combining within-formant spectral measurements from various smaller segments is novel. To our knowledge, no previous study has explored these features together in a single model. The current study presents a new model based on spectral features (spectral moments, amplitudes, bandwidths, and spectral peaks) taken from the midpoints of the first four vowel formants and explores their function in speaker comparison.

The study includes within its focus an under-researched formant, F4. While F1 - F3 and their significance in speaker discrimination have been researched repeatedly (McDougall, 2004, 2006), but most studies have excluded F4. Zhang et al. (2013) suggest that F4 has been under-researched mainly because many forensic recordings are of telephone-transmitted speech, and until very recently this had an upper-frequency limit of around 3400 Hz for landlines. With recent advances in mobile and social media telecommunications, these limitations have been reduced. This gives rise to a need for the possibility of including values associated with F4 in the forensic casework (Cao & Dellwo, 2019). , The following sections provide a brief overview of spectral measures and moments.

8.4.1 Spectral measures

The role of within-formant spectral measures, such as formant amplitudes, formant bandwidths, and spectral peaks, has been investigated by various researchers (e.g., Alam et al., 2015; Gonzalez-Rodriguez, 2011; Hillenbrand et al., 2006; Jacewicz, 2005). Recent findings by Suthar and French (2023b) suggest that analysing these measures in conjunction with formant centre frequency values significantly improves the predictability of the model by up to 200 per cent. Formant amplitude, in particular, exhibits a strong correlation with formant centre frequencies (Kent & Read, 2002), indicating its potential to carry speaker-specific information akin to formant centre frequencies. The inclusion of formant amplitudes in models

where centre formant frequency peaks are difficult to extract or including them with formant centre frequencies (after eliminating the associated features), may be valuable for forensic case studies. The study also revealed that amplitude outperformed spectral peak and bandwidth measures.

Formant bandwidth analysis showed that vowel perception is rarely affected by their respective bandwidths, but any drastic reduction or increase in the bandwidth can make vowels sound artificial (Hawks & Miller, 1995; Kent & Read, 2002). For formant bandwidths, other research has proposed that they have been associated with the identification of the sex of the speaker since analysing bandwidth for female voices provided very different results from that of males, and females have slightly higher bandwidths (Kent & Read, 2002). Considering all these factors, Suthar and French (2023b) looked at the impact of analysing formant bandwidths in the context of forensic speaker comparison for female speakers. Their study showed that assessing formant bandwidths (both spectral and LPC) with formant frequencies increases speaker classification rates above the chance level up to 6.9 times over the chance level, i.e., 590 per cent. The same study also showed that when combining these bandwidth measurements in conjunction with certain features (formant amplitude, spectral peaks), some combinations can increase the performance up to 12 times above the chance level (i.e., 1000 per cent above chance).

Given the individual (e.g., Alam et al., 2015; Gonzalez-Rodriguez, 2011; Hillenbrand et al., 2006; Jacewicz, 2005) and combined (Suthar & French, 2023b), the significance of amplitude, bandwidth, and spectral peaks in forensic speaker comparison studies and the observed potential enhancement through their interaction with formant centre frequencies after removing the correlated features, this study aims to investigate their collective role in forensic speaker discrimination. In the subsequent section, we provide a brief introduction to spectral moments, another set of acoustic features used in the study.

8.4.2 Spectral moments

Spectral moments are numerical distributions of acoustic energy. [19] Spectral moment analysis , just like spectral measure analysis, has been investigated in research motivated by the needs

[19] For further details on spectral moments please see Suthar & French (2023a).

of forensic speaker comparison forensic speaker comparison (Eriksson, Cepeda, Rodman, McAllister, et al., 2004; Suthar & French, 2023a; Weingartová & Volín, 2013).

Our study explores four primary spectral moments (Nittrouer, 1995). These are the centre of gravity (COG), standard deviation (SD), skewness (Skew), and kurtosis (Kurt). Weingartová and Volín (2013) discussed the significance of employing shorter segments for forensic cases instead of long-term spectra, as the latter is affected by various factors such as the length of the speech utterance or by the content analysed, thus making it incomparable and less reliable in forensic cases. Their work proposed using smaller chunks or short-term spectra for speaker comparison work to compute the spectral slope. Building on this, Suthar and French (2023a) analysed the first four primary spectral moments extracted from smaller spectral chunks, i.e., formant centre frequencies of vowels. Studies have suggested the discriminatory power of acoustic-phonetic measurements decreases when analysed in isolation (Cavalcanti et al., 2023; Hughes, 2013; Künzel, 1997), i.e., adding more features into a single model provides a higher speaker discriminatory potential than a system with only limited number of features. The study proposes combining acoustic-phonetic measurements for improved speaker discrimination and suggests that shorter segments of speech are more reliable for forensic cases. The study showed a drastic increase in the correct prediction rate by up to 7.8 times above the chance level for certain spectral moment combinations. The study also mentioned that COG was the best-performing feature, which is to be expected as, just as amplitude, COG is also highly correlated with formant centre frequencies.

We propose here to merge the models set out in Suthar and French (2023a) and (2023b) to create a new model for forensic speaker discrimination. The acoustic measurements set out in both studies can be easily extracted with the help of a Praat script and replicated when needed, essentially making it a very significant addition to forensic casework. To achieve this, the paper will look at the probability of correct participant classification based on the model with the help of linear discriminant analysis (LDA). For uniformity, the term spectral feature will be used to represent the combination of both spectral moments and spectral measures. As demonstrated in Suthar and French (2023a, 2023b), factors such as vowels, varieties, and speech style (mode of data elicitation) affect the spectral moments and measure values. The papers also suggested that there are certain vowels, varieties, and speech styles where the LDA performs better for the selected features than the others. Considering both premises, the questions devised for the present study are as follows:

1. Which feature clusters from F1–F4 extracted from around mid-point in a vowel have the highest speaker discriminatory power: bandwidth, within-formant skew, within-formant kurtosis of energy, formant amplitude, relative amplitude centre of gravity, standard deviation, spectral peak?
2. Does combining within-formant spectral moments (centre of gravity, standard deviation, skewness, and kurtosis) and spectral measures (formant amplitude, formant bandwidths, and spectral peaks) improve the accuracy of speaker classification? If so:
    2.1. Which spectral feature combination has the greatest speaker discriminatory value?
    2.2. Which vowels and vowel subsets have the greatest discriminatory value when subjected to spectral feature analysis?
    2.3. Do spectral features or feature combinations perform better for some modes of data elicitation than others as speaker discriminatory features?
    2.4. Does the speaker discriminatory power of spectral feature analysis work better for some varieties than others?

8.5 Language

The study is based on speech samples in Marwari, an Indo-Aryan language spoken in India. Participants were 45 female monolingual speakers representing three different caste dialects: Bishnoi, Jaat, and Brahmin. 15 speakers from each caste variety were recruited from the Bikaner district of Rajasthan (India). A brief description of the three caste dialects is provided:

Brahmin: Traditionally associated with priests and religious scholars in Hindu caste systems.

Jaat: This caste primarily comprises farmers and warriors.

Bishnoi: A relatively new caste, founded in 1485 by Guru Jambheshwarji.

It is noted that phonetic differences exist among the three caste dialects, as depicted in Figure 8.1. The figure depicts the vowel space occupied by the vowels chosen for the current study in three distinct varieties of the Marwari language.

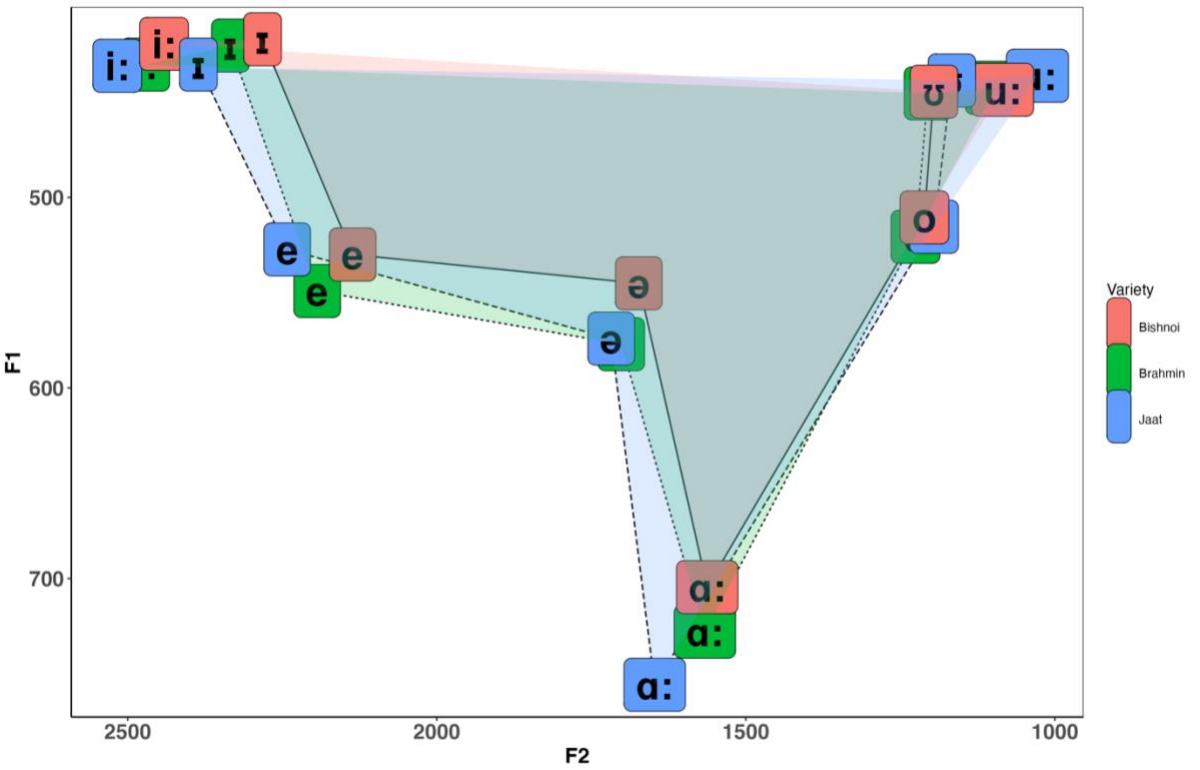


Figure 8.1 The vowel space chart of three varieties of Marwari for all three data types together.(Green = Brahmin, Blue = Jaat and Red = Bishnoi)

8.6 Methodology

Data processing started by gain-normalising the peaks at 2 dBFS (decibel full scale) with the help of Soundforge 2.0. This step was followed by isolating the target sounds on Praat and discarding any noisy recordings. The extraction of formant centre frequencies was initially done with the help of a Praat script and later corrected manually to ensure the accuracy of the values. Following this, a detailed extraction of spectral features was conducted with the help of a Praat script (Harrison, 2021). To make feature extraction more accurate, eight different settings were selected. The settings provided manual control over deciding how to extract the within-formant features from the centre frequencies. The feature extraction relied on the amplitude drop (+/-3 dB versus +/-1 dB) and different smoothing settings for the harmonics. The script first smoothed the harmonics to make the formant centre frequencies more visible, making it easier to pick them up automatically. The script was designed to extract formant centre frequencies based on the manually extracted formant centre frequency values, i.e., the script automatically identified the peaks closest to the previously acquired formant centre

frequency data and chose the nearest possible values. At this stage, +/- 3 dB amplitude was selected after visual examinations of each extracted token.

8.6.1 Spectral measure estimation

Formant amplitude is based on the sound pressure level of the nearest formant peak for individual formant centre frequencies (extracted sound pressure levels as A1 - A4 for F1 - F4). As one of the most used methods of analysing amplitude is to analyse relative amplitudes (RA), amplitude analysis was also conducted for the RAs between pairs of formant peaks. The amplitude difference between the two peaks provided six different RA measures: Amplitude of F2-Amplitude of F1 (A2-A1), Amplitude of F3-Amplitude of F1 (A3-A1), Amplitude of F4-Amplitude of F1 (A4-A1), Amplitude of F3-Amplitude of F2 (A3-A2), Amplitude of F4-Amplitude of F2 (A4-A2) and Amplitude of F4-Amplitude of F3 (A4-A3).

Spectral bandwidth (SB) was extracted by calculating the difference between +/- 3dB upper and lower frequency for formant peaks.

$$fnSB = fnUpperFreq - fnLowerFreq$$

$$(fnLowerFreq = fnSP - (fnAnalysis\ band/2)$$

$$fnUpperFreq = fnSP + (fnAnalysis\ band/2))^{20}$$

The script also extracted the LPC bandwidth (LB) with the help of Praat for further analysis. Spectral Peaks (SP) were extracted from the nearest maximum from the manually extracted formant centre frequencies.

8.6.2 Spectral moment estimation

Spectral moments were extracted automatically using Praat's default settings. The script automatically measured the centre of gravity (COG), standard deviation (SD), kurtosis (Kurt), and skewness (Skew) within Praat's spectral slice function. These measurements were extracted from the pre-defined spectral slice extracted by the script at a drop of +/-3 dB amplitude from either side of the formant peak.

[20] The analysis bands were modified by trial and error to find the optimum settings. The final bands selected with the least number of errors were F1 = 300, F2 = 500, F3 = 600, F4 = 700.

8.6.3 Statistical Analysis

An ANOVA conducted in R for wordlist (1), story (2), and conversation (3) data for each feature showed significant inter-variety and inter-vowel differences. The next step was assessing each feature's speaker discrimination value. To test the significance of each feature for different varieties, and vowel effect, two models (with participants being random variables) were created. The first model treated the variety and vowel alone as factors, and the second model had vowel and variety interaction as factors. These models were tested for three different modes of data elicitation. Both vowel and variety as factors played a significant role, and their interaction also presented p-values <.05 (i.e., significant). The next step was to check for the assumptions posited by linear discriminant analysis (LDA), i.e., that the data is normally distributed, and each class has equal covariance. The data was z-transformed for normalisation, and 'box-M' tests were conducted to check for covariances. The correlations were also tested as an additional test to check if no two features were highly correlated.

Figure 8.2 demonstrates that COG was positively correlated with their respective spectral peaks (SP) and formant centre frequencies (F1 - F4). Additionally, the figure also shows that the amplitudes had significant correlations among themselves and, in some instances, with COG. The correlational analysis helped select features that could go into the same model. LDA was conducted on the features extracted from every variety to predict these features' classification rate (CR) in identifying individual participants.

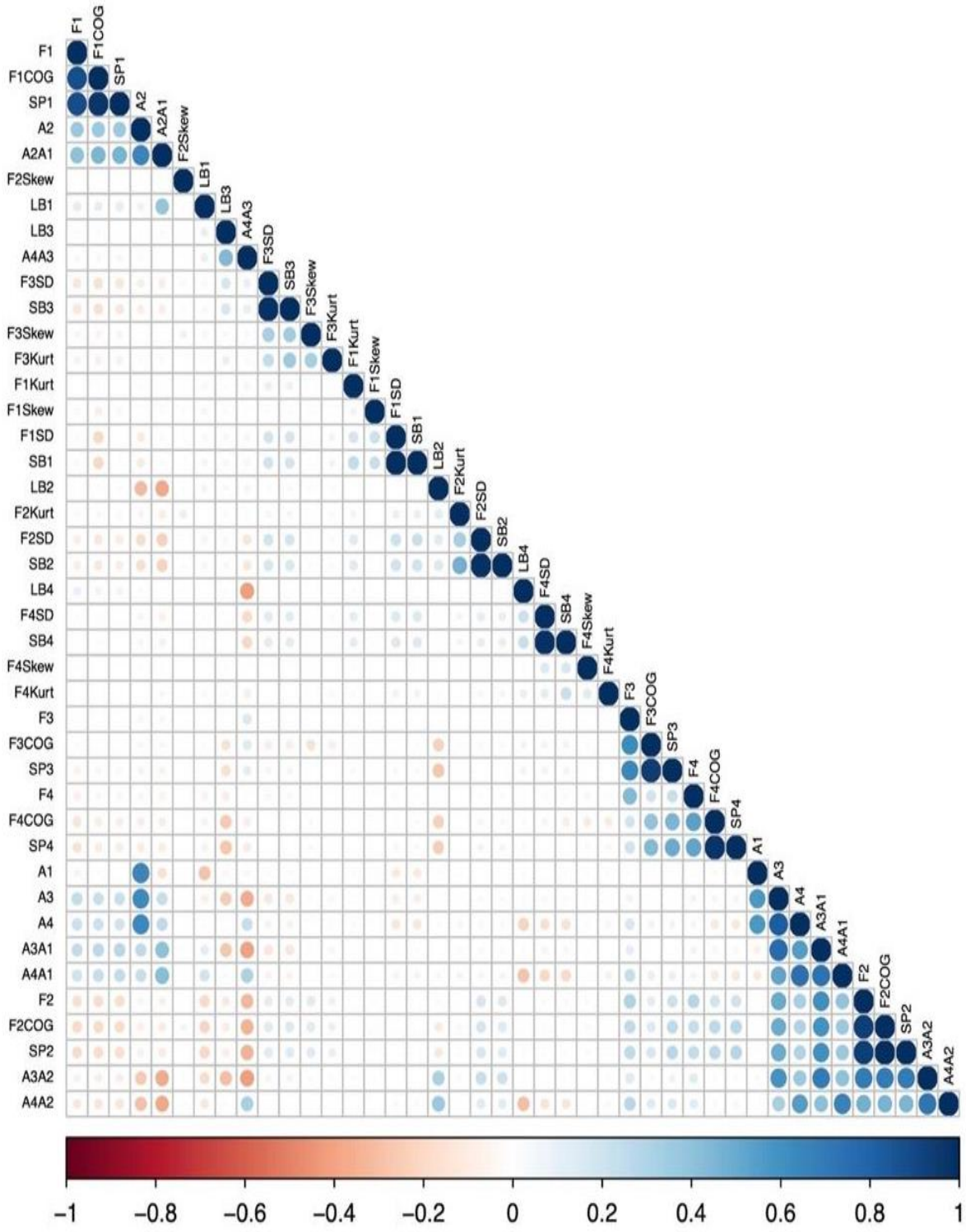


Figure 8.2 Correlation between features with formant centre frequencies. Highly correlated variables are marked by blue and least correlated variables are marked by orange/red.

8.7 Results

8.7.1 Within-formant features analysis for individual and combination of features

The first step was to check the individual feature's performance. When tested alone, the mean CRs ranged from 2.9 times over the chance level to up to 5 times over the chance level. The features that performed the best were formant amplitudes extracted from F3 and F4. These features demonstrated the highest discriminatory power.

Following this, the next step was to verify if including these features with formant centre frequency values increased the performance of the model (after testing and removing the correlated features). Figure 8.3 combines the results from two studies, Suthar and French (2023a) and (2023b), to analyse the impact of adding spectral features with centre formant frequencies on correct classification rates (CR) for three different data types, wordlist, story and conversation.

The results indicate that simply adding one measure with all four formant midpoint values led to a significant increase of over six times above the chance in CRs for Wordlist data. In the figure, it is visible that for Brahmin, the average CR increase was 6.09 times over the chance level, which was a substantial increase from the model where only the first four formants were included. The CR values for varieties 1 and 3 had a clear distinction between them for all three data types, which was not that clearly visible in Jaat. Wordlist in general had higher classification rates than the other two.

However, Figure 8.3 does not account for the collinearity of the features; for example, COG values were tested with centre formant frequencies, which were found to be highly collinear (as shown in Figure 8.2). To address this, a new analysis was conducted where only non-collinear features were put together in the model. At this step, multiple models were created. Through these models' multiple combinations, starting from one at a time to four. The number was limited to four as LDA has an n-1 limitation on the number of variables that can occur in an analysis. With four features and four centre formant frequencies, the total number of variables reached up to eight. The number of tokens for the present study was limited to ten, and in most cases, during the extraction process, some tokens were removed, leaving only nine tokens per vowel for the analysis. The new models also looked at different modes of data elicitations and varieties.

The new model has been summarised in Figure 8.4, which has been divided into three parts, representing three varieties. It also shows three different modes of data elicitation through different coloured hexagons. The figure is arranged by adding four features with formant frequencies to one (top to bottom). The figure shows that type one was again the best-performing mode, followed by type two. A noteworthy point shown in the figure is that the average CR decreased as more variables were included in the model. This suggests that adding more features led to some loss of CR performance. This can be clearly observed by the line depicting the averages for each model combined.

The next step of the analysis included identifying the best-performing measures from each data type and variety, and assessing if adding these measures in various combinations can improve the model. Based on these conditions, nine new models were created, each consisting of eight best-performing non-collinear features. The combination models were based on high-performing individual features, including both lower formants (F1, F2) and higher formants (F3, F4). The hypothesis was that performance would improve drastically once the combinations of features were tested. However, this was not true for the eight best-performing features. While some models provided a ten-fold increase in predicting speakers with the help of LDA, the average increase was limited to five times (see Table 8.1). The results of this analysis are very similar to those of adding four variables to formant measures presented in Figure 8.4, i.e., the CRs do increase for as much as 5.3 times above the chance level, but this increase can also be obtained by just adding two/three variables to the model. This suggests that the optimum number of measures for the current data is five measures to seven measures in the same model.

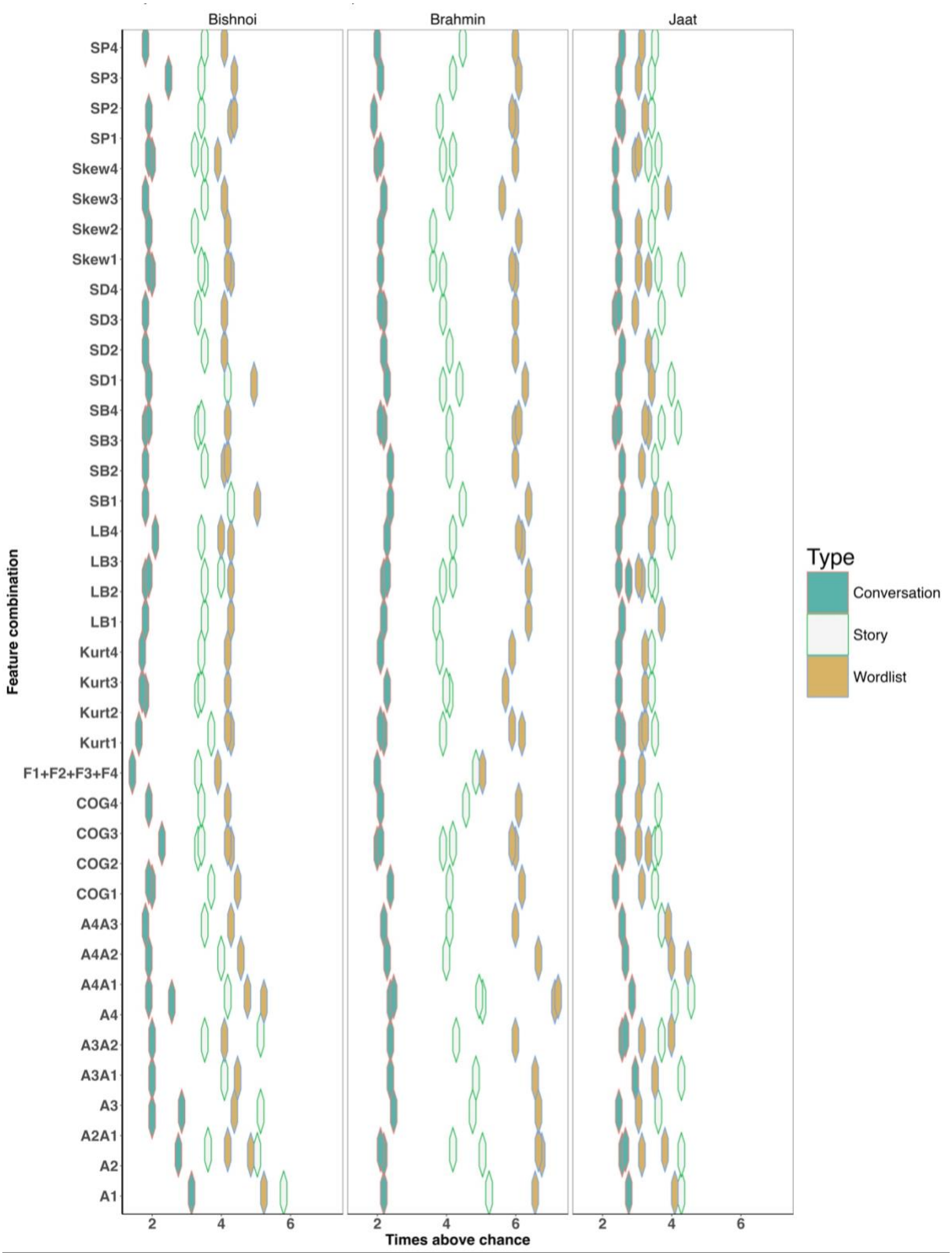


Figure 8.3 Times above chance classification rates for individual features combined with F1-F4.

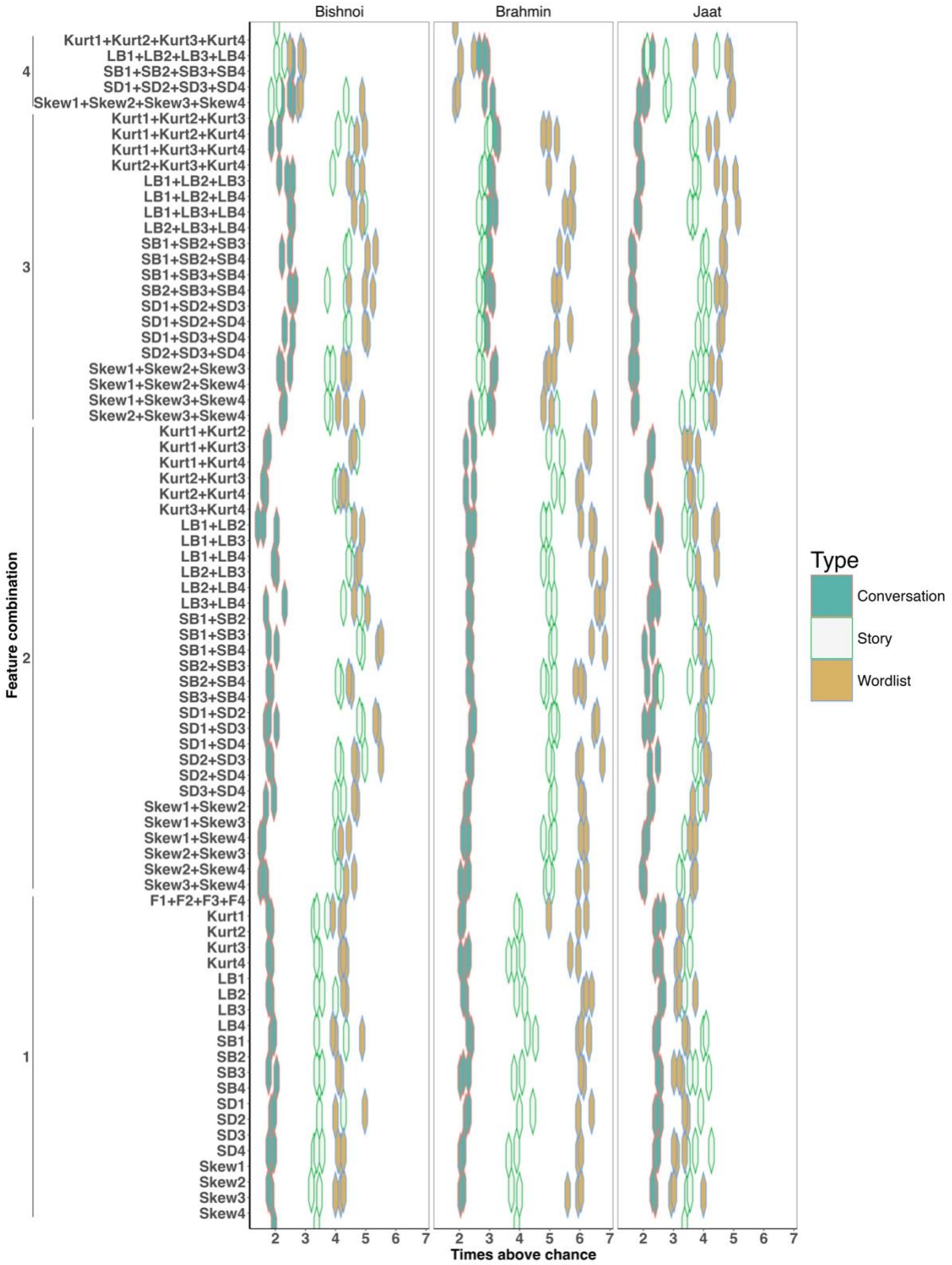


Figure 8.4 Combination of features with centre formant frequencies.

Table 8.1 Performance of models (Word. = Wordlist, Conv.= Conversation)

		Brahmin			Jaat			Bishnoi		
		Word.	Story	Conv.	Word.	Story	Conv.	Word.	Story	Conv.
M1	A4A1+A1+A4+SB1+LB1+LB2+ LB3+F1KURT	1.8	-	2.3	4.2	2.3	2.1	2.3	1.3	2.9
M2	A4A1+A4A3+A1+A2+LB1+LB4+SB1+F4SD	2.1	-	2.2	5.2	2.6	2.3	2.5	2.1	2.7
M3	A4A2+A3+SB1+SB2+LB4+F1COG+F1SD+F2SD	2.2	-	2.8	4.5	2.2	2.1	1.9	2.1	2.5
M4	A2A1+A3A1+A4A1+A4A3+A1+LB4+SB1+F3SKEW	2.1	-	1.9	5.1	2.6	2.4	2.4	1.8	2.6
M5	A2A1+A3A1+A4A1+A4A3+A1+A2++SB1+F4SD	1.9	-	2.2	4.7	2.7	2.2	2.4	1.8	2.9
M6	A2A1+A3A1+A1+LB1+LB2+SP4+F3COG+F1KURT	2.3	-	2.3	5.1	2.6	2.1	2.6	1.6	2.6
M7	A4A1+A4A3+A1+SB1+LB1+SP3+F1COG+F3COG	2.1	-	2.7	5.3	3.1	2.3	2.4	2.1	2.8
M8	A2A1+A4A1+A1+SB1+LB2+LB3+F1COG+F1KURT	2.1	-	2.4	4.4	2.5	2.1	2.2	1.8	2.9
M9	A2A1+A3A1+A1+SP3+LB3+LB4+F1COG+F1SD	2.3	-	2.3	4.8	2.9	2.2	2.3	2.4	3.0

8.7.2 Within-formant feature analysis for different vowels

The performance of vowels for individual features varied widely, ranging from 0.25 times greater than the chance level to 4.5 times. However, the higher CRs were limited to specific vowels. In other words, certain vowels consistently outperform others in terms of correct CRs. Nevertheless, there was a lack of a clear pattern, as sometimes one vowel performed well for certain features but not others. For example, the average performance of vowels based on different types was as follows:

Wordlist: /a:/ > /i:/ > /e/ > /o/ > /ɪ/ > /ʊ/ > /u:/ > /ə/

Story: /i:/ > /o/ > /ɪ/ > /ʊ/ > /e/ > /a:/ > /ə/ > /u:/

Conversation: /ʊ/ > /i:/ > /a:/ > /e/ > /ɪ/ > /o/ > /ə/ > /u:/

This could be attributed to the limited number of tokens available once the data was divided based on vowel categories. The number of tokens further declined for different modes of data collection. This reduced the number of dimensions that could be put into an LDA model. Nonetheless, the analysis still provided improved CRs when features were added with centre formant frequencies. Figure 8.5 and Figure 8.6 show the results of vowel analysis for vowels /a:/, /e/, /i:/, /u:/, and for vowels /o/, /ʊ/, /ɪ/, /ə/, for the Type 1 data. The figures also show the average CRs for each vowel (presented by a dashed line). The average CRs for these vowels were at least two times higher than the chance level. Vowel /a:/, on average, performed better than other vowels. However, the highest CR was observed for the vowel /o/, which was over 5.3 times higher than the chance level. The analysis further revealed that the best-performing feature had a clear and observable pattern for all three data types, i.e., amplitudes performed better for every vowel. The second-best performing features were SPs and the COGs. For Story

data, the average CRs decreased compared to Type 1[21]. However, similar to Type 1, the amplitude and spectral peaks were again the best-performing features, followed by the COG values extracted from centre formants.

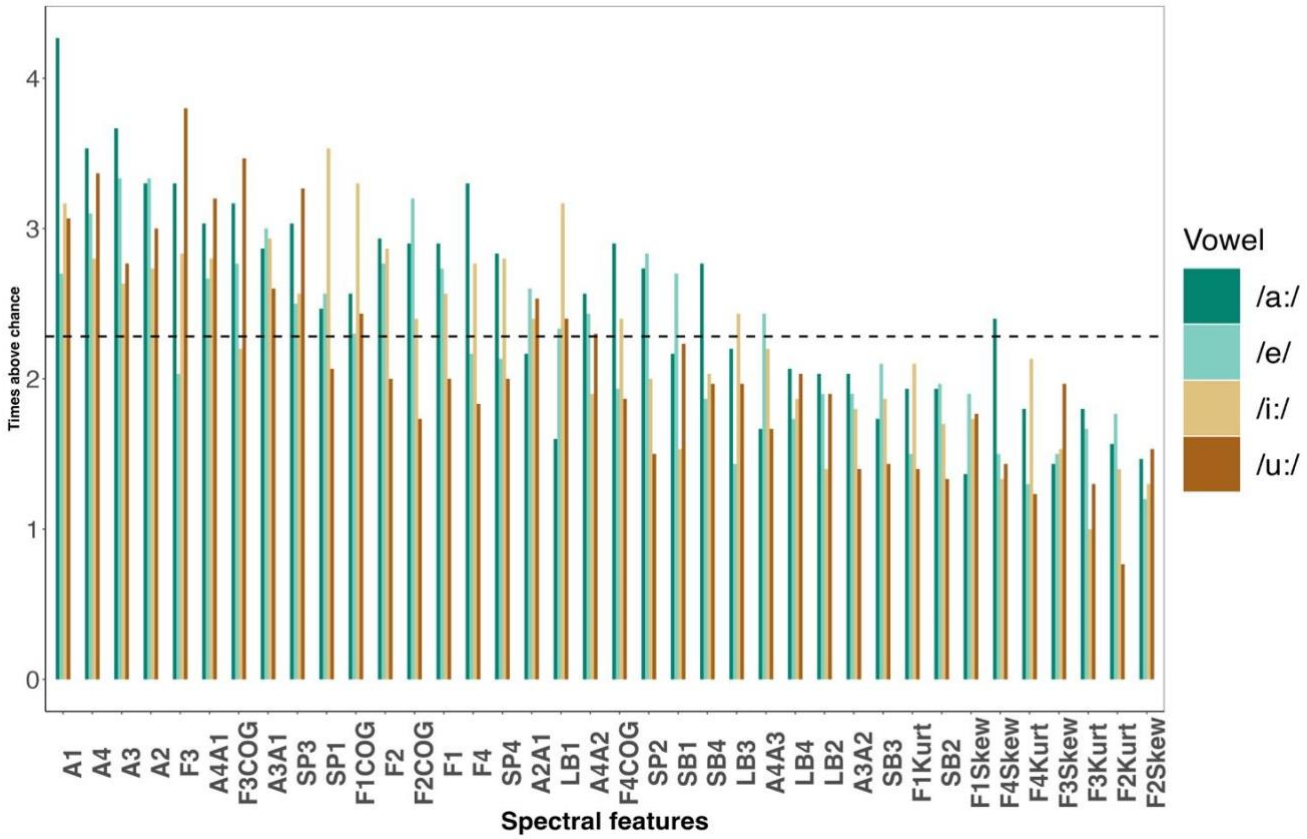


Figure 8.5 Individual performances of vowels /a:/, /e/, /i:/, /u:/ of for wordlist data.

[21] Graphs for individual vowels for story and conversation data are provided in the appendix.

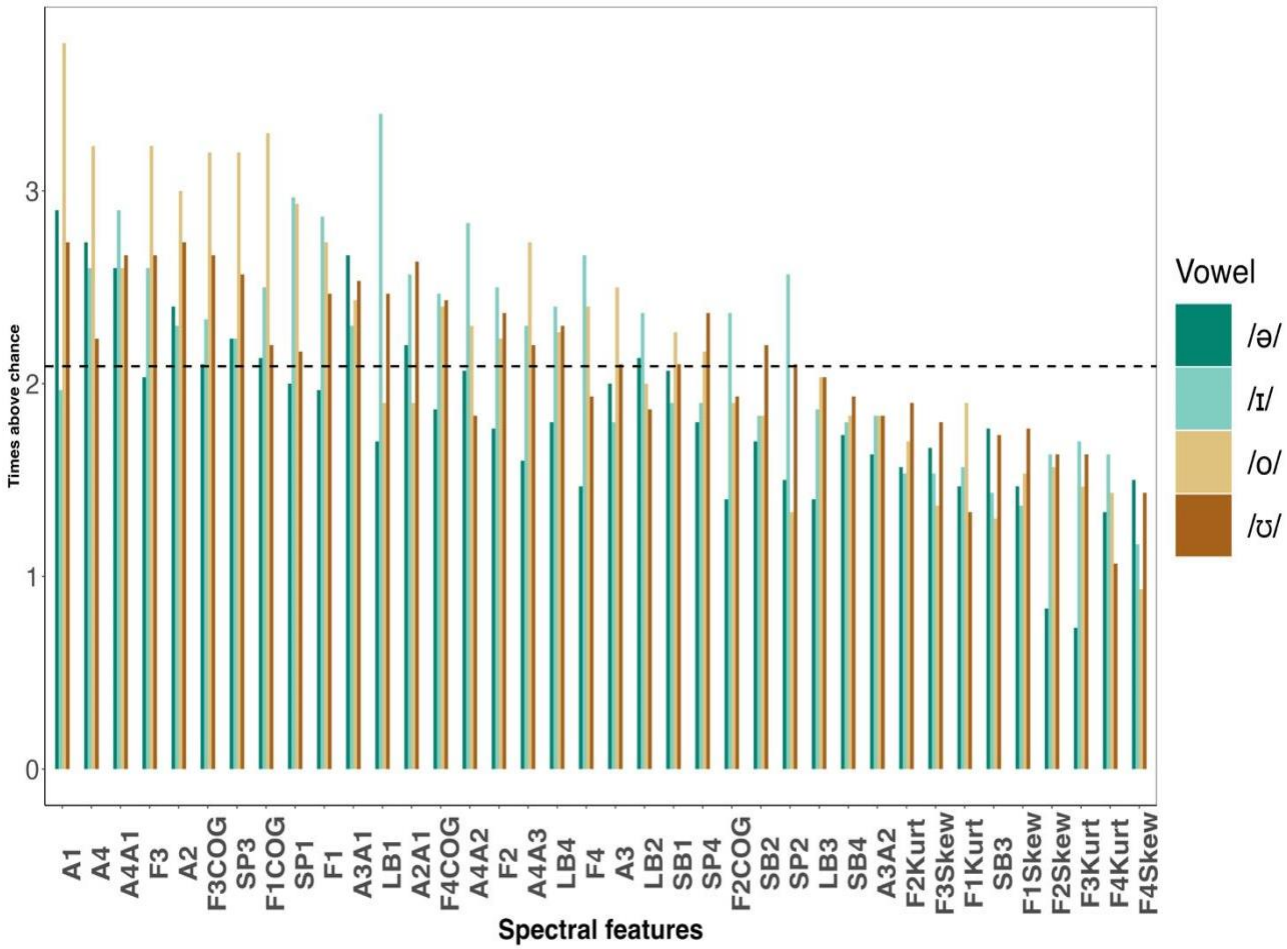


Figure 8.6 Individual performances of vowels /o/, /ʊ/, /ɪ/, /ə/ of for wordlist data (see Appendix A for the performance of vowels for other types of data).

Overall, the results showed that the choice of vowel significantly influenced the performance of various features used in speaker classification. Certain vowels consistently led to higher CRs, suggesting that their acoustic characteristics were more distinct and easier to differentiate. These findings highlight the importance of considering choice of vowels as an important factor in speaker classification studies based on within-formant features.

The combination models for vowels were tested for the nine best-performing models. Among these models, Model 9 consistently outperformed the others, showing higher average CRs across all three modes of data elicitation. Interestingly, the best-performing subset of vowels was the one composed of closed vowels. This model achieved CRs as high as 15.2 times above the chance level. This indicates that closed vowels, in general, demonstrated superior performance compared to the other vowel categories.

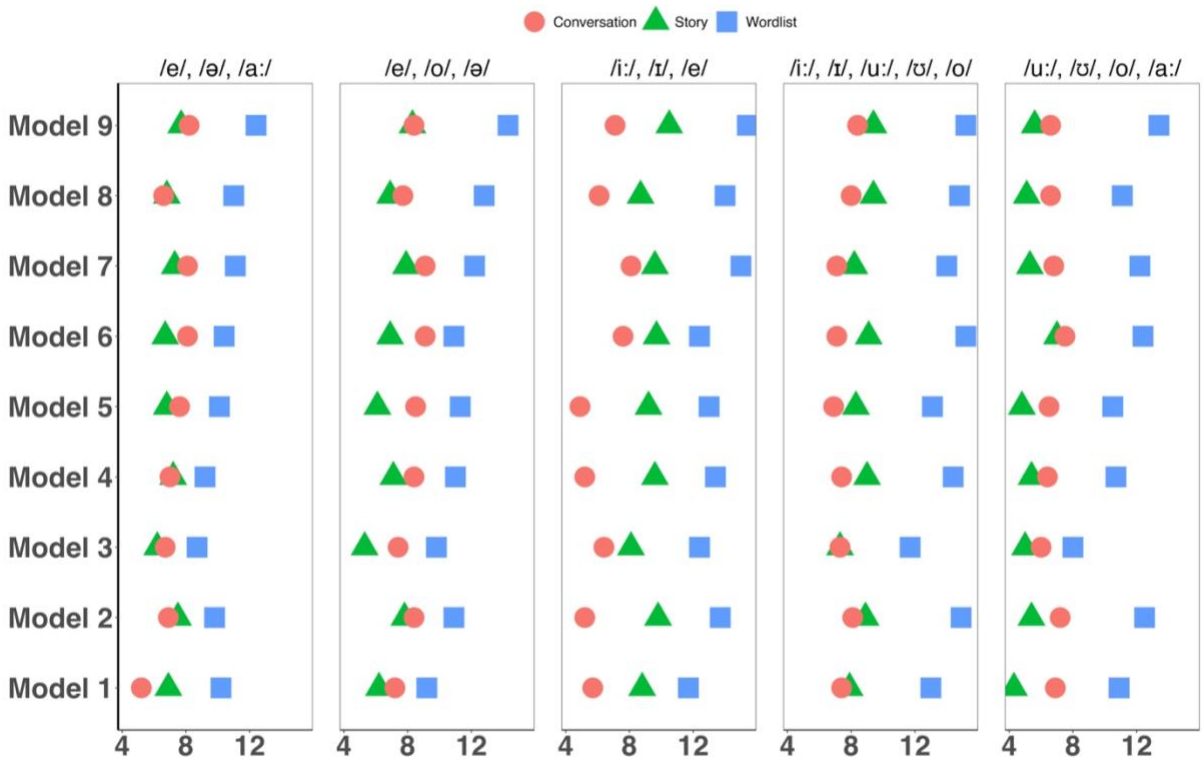


Figure 8.7 Vowel subset analysis for different models. The x-axis represents the CR values for each vowel subset. Y-axis presents the models in descending order. Different types of data is presented with different shapes on the graph with the circle showing conversation, triangle representing the story value and a square showing the wordlist value.

8.7.3 Within-formant feature analysis for different modes of data elicitation

As mentioned in sections 8.7.1 and 8.7.2, Wordlist data performs exceptionally better than the other two for both single and combinations of features. The average for Wordlist data was 4.6 times the chance for single features, whereas for Conversation, it was 2.2. Story was always the second-best performing data elicitation mode, with an average CR of 3.8 for single features.

For combination models, the same pattern was repeated, suggesting spectral feature analysis works best on wordlist data.

8.7.4 Within-formant feature analysis for different varieties

Brahmin consistently performed better than the other two varieties, but the differences in the average CRs for each feature were minimal. For example, the CR differences for feature A4

across all three varieties were 4.8, 3.5, and 4.2, indicating relatively small variations in performance among the varieties.

The performance of the three varieties for the three different types of data was notably different. The average CR for Wordlist data in Brahmin was six times over the chance level, which was four times higher than the average CR for Conversation data in the same variety. This suggests that Brahmin performed significantly better with Wordlist data compared to Conversation.

Further, the average CR values for different varieties were also distinct. For Wordlist data, the average CRs for Brahmin, Jaat, and Bishnoi were 6, 4, and 2, respectively. For Story data, the average CRs were 3, 3, and 2 for Brahmin, Jaat, and Bishnoi, respectively. Lastly, for Conversation data, the average CRs were 4, 3, and 2 for Brahmin, Jaat, and Bishnoi, respectively.

Looking at the individual features, amplitude consistently performed better for each variety, with the highest CR observed for Brahmin (A4), at 5 times above the chance level; for Jaat, 4 times (A4); and 4.6 times above the chance level for Bishnoi (A1). Delta values between formants also showed strong performance, with the highest CR observed for Brahmin (A4A1) at 4.8 times, for Jaat (A4A1) at 3.9 times, and for Bishnoi at 3.6 times over the chance level. These features were followed by LPC bandwidth and COG CRs.

Overall, each feature improved the CRs by a minimum of 2.5 times compared to the chance level.

For the combination of features, Brahmin again performed consistently better than the other two varieties, and the variation between the CRs was minimal. This indicates that the different combinations of features did not have a significant impact on the overall performance of each variety.

Additionally, the average performance of CRs was found to be dependent on the number of variables included in the model. Specifically, models with 6 or 7 features provided higher CRs for each variety compared to models with 8 features. This suggests that there may be an optimal number of features for each variety beyond which adding more features would not lead to substantial improvements in classification accuracy. The same results were repeated for the best-performing models, as presented in Table 8.1.

8.8 Summary

We began with the following research questions, which, for ease of reference, are reproduced here:

1. **Which feature clusters from the F1 – F4 centre formant frequency range have the highest speaker discriminatory power: formant bandwidth, within-formant skew, within-formant kurtosis of energy, formant amplitude, relative amplitude, centre of gravity, standard deviation, spectral peak?**

A.1. The best-performing features were formant amplitudes.

2. **Does combining within-formant spectral moments (centre of gravity, standard deviation, skewness, and kurtosis) and spectral measures (formant amplitude, formant bandwidths, and spectral peaks) together improve the accuracy of speaker classification?**

A.2 Combining formant spectral moments and measures does increase the accuracy of the speaker classification model.

If so:

2.1 Which spectral feature combination has the greatest speaker discriminatory value?

A.2.1. The best-performing speaker classification models were ones that were created with the following combination providing the highest CR observed, i.e., 15.4 times over the chance level.

A2A1+A3A1+A1+SP3+LB3+LB4+F1COG+F1SD

(relative amplitude values of formant amplitude differences between second formant amplitude-first formant amplitude and third formant amplitude-first formant amplitude combined with third formants' spectral peak, LPC bandwidth for third and fourth formant, first formant's centre of gravity and standard deviation)

When combined with centre formant frequencies, the optimal number of features that should be added to the model was up to 6/7. After this, the performance of the model decreased.

2.2 Which vowels and vowel subsets have a greater discriminatory value for spectral feature analysis?

A.2.3. The best-performing vowel was the vowel /a:/, and the best-performing vowel subset was the one created with close vowels.

2.3 Do spectral features or feature combinations perform better for some modes of data elicitation and associated speech styles than others as speaker discriminatory features?

A.2.4. Spectral feature analysis performed best for Wordlist (wordlist) data.

2.4 Is the speaker discriminatory power of spectral features better for some varieties than others?

A.2.5. Brahmin performed better than the other two varieties, although the difference between the CRs was minimal.

8.9 Discussion

The paper examines the significance of including acoustic features of eight different vowels in a speaker comparison model, analysing 42 different acoustic characteristics. Among these features, three stood out as the best performers individually: amplitudes, spectral peaks, and COG for all four formants. One of the reasons for the higher performance of formant amplitudes and spectral peaks could be that, as perceptual features, they are influenced by the interaction of formant peaks during vowel production, and just like formant, these features can be affected by changes in the vocal effort during speech (Kent & Read, 2002). Their paper also mentions that formant patterns and their interaction during vowel production can lead to higher amplitude values. This could explain why amplitude values are directly related to the formants centre frequencies and, like these formant centre frequencies, provide more robust results for individual speaker classifications because of vocal effort changes.

Results also suggested that the performance of LBs was better than that of SBs. This could be explained by the fact that SBs, as mentioned in Section 8.6.1, were extracted from spectral peaks, thus being more prone to errors than LBs (McCandless, 1974). The errors for SBs can be accounted for by the fact that spurious peaks derived from two closely spaced formant peaks

can result in errors (McCandless, 1974). The same could explain the performance of COG, which was as good as its respective formant centre frequencies. These findings are in line with those of (Chistovich & Lublinskaya, 1979). The paper hypothesises that COG could be a better indicator of individual speaker discriminant for female voices than centre formant frequencies. Formant peak extraction for female voices is relatively more difficult than for male voices because of the higher F0 of female speech and, as a result, a lower number of harmonics around the lower formant (Künzel, 2001). COG, unlike formant peaks, relies on the entire energy distribution of the within-formant spectral slice, thus providing a more robust representation of the vowel sounds in the high-pitched recordings of female participants, even in the presence of combined formant centre frequency peaks.

The variability of formant measurements in the eight best-performing measures was from both lower and higher formants, suggesting that though higher formants have been proven significant in speaker classifications (McDougall, 2004), spectral features extracted from the first four formants do not follow these patterns.

For single vowels, vowel /a:/ performed higher than the rest, but for vowel subsets, close vowels performed better than any other category of the vowel subsets. This could be because vowel /a:/ provided the optimum settings for spectral feature extractions in comparison with others. The same could well be true for close vowels as a set. As for close vowels, the constrictions of the vocal tract result in higher formant frequencies for shorter durations compared to open vowels. This could be further explained by the fact that for close vowels, the constriction of the vocal tract results in shorter vocal tract openings. As a result, close vowels tend to have higher resonant frequencies and thus have higher energy concentrations available for spectral feature extraction and analysis than their open counterparts. For spectral features extracted from these shorter segments from higher frequencies, they might provide more information in comparison to that of the open vowel set. The open-close distinction here also bears on the significance of F1, showing that as closed vowels were the ones performing better, higher frequencies of F1 provided more robust speaker information over the lower F1 frequencies provided by open vowels (Weingartová & Volín, 2013).

Brahmin performed better than the other two varieties, but as mentioned earlier, this difference was minimal. As every other aspect of the analysis was controlled, the difference in performance can be attributed to the recording environment. Brahmin was recorded in an urban

home with smaller room sizes and less echo, while the other two varieties were recorded in rural areas with larger room sizes and more echo. Despite these differences in recording environments, the overall results show that spectral feature extraction can still be used effectively in scenarios where recordings were extracted from different conditions. It also suggests that spectral feature analysis is robust enough to handle variations in the recording environment and can still provide meaningful results, even when the recordings are made in different acoustic settings. To test this proposition, various tests of the same features in a range of different settings would be needed.

8.10 Limitations

Despite certain features demonstrating relatively high performance, the overall increase in CRs with most features was relatively low. One possible reason for this could be the influence of several factors on the study. For instance, once the data was classified based on vowels, the number of analysed tokens was drastically reduced, which might have impacted the results. Additionally, data was acquired from participants' individual houses to achieve forensically realistic data. This could have impacted the various spectral parameters, such as skewness and kurtosis.

Studies have pointed out that formant-related analysis can be challenging for various reasons, such as the fact that any automatic formant extraction would require human supervision and that extracting formants manually can be time-consuming (Duckworth et al., 2011; Kinoshita et al., 2022; Zhang et al., 2013).

As mentioned in Section 8.7, more controlled recording settings are needed to further verify the results.

8.11 Implications

The study recommends that in manual speaker discriminant analysis for both forensic speaker comparison casework and academic research, spectral feature analysis, including spectral moments, should be included in addition to commonly used acoustic features such as formant centre frequencies, fundamental frequency, and voice quality analysis. The study also suggests that, for more robust inter-speaker variations, closed vowels should be analysed along with spectral feature analysis. The closed vowel values extracted from the steady state of stressed

vowels can be valuable additional tools alongside other standard acoustic features for accurately identifying individual speakers. One positive benefit of the fieldwork and circumstantial limitations associated with women in rural India, where there was little to no chance of getting them to attend any kind of recording session, even if it was available, was that the non-matched acoustic environments approached those encountered in real forensic speaker comparison casework.

9 Chapter 9: Conclusion

Chapter 9 provides a conclusion for the study. It begins with a summary of the results reported in Articles 1 - 3 and then moves on to the study's general conclusion. The chapter finishes with the current study's shortcomings and implications.

9.1 Summary of the Articles

9.1.1 Article 1

Article 1 assessed the efficacy of Spectral moment analysis (SMA) in identifying individual speakers based on spectral moments (centre of gravity, standard deviation, skewness and kurtosis) from vowel formants and explored its effectiveness and application to various speech styles.

1. **Factors affecting spectral moment values during extraction:**

   The results indicated that the values of SMA were affected by vowel and mode of data elicitation.

2. **Effective Spectral Moments:**

   Among the spectral moments, the Centre of Gravity (m_1) emerged as the most effective in contributing to speaker discrimination.

3. **Factors affecting the discriminant values of SMA:**

   SMA's effectiveness in distinguishing speakers depends on vowel choice, but its effectiveness varies with speaking style, indicating no clear pattern with varying discriminant values.

   The study revealed that SMA's performance was most effective when applied to wordlist data, implying that the quality, steady state of vowel extraction and characteristics of the speech data significantly influenced its outcomes.

   SMA demonstrated consistent performance across all three Marwari varieties, indicating its effectiveness in speaker discrimination across different linguistic groups.

In conclusion, the study emphasizes the practical value of SMA in speaker discrimination tasks, especially with vowel formants, and its effectiveness depends on spectral moment choice, vowel and variety choice, and data elicitation mode. Further research on SMA could help assess its generalizability across languages and models, and if it can be used as a valuable tool for speaker discrimination in real-world scenarios.

9.1.2 Article 2

Article 2 investigated the impact of various factors on the spectral measure (formant amplitude, formant bandwidths and spectral peaks) and their effectiveness in distinguishing individual speakers in acoustic analysis. Major findings and conclusions were:

1. **Factors affecting spectral measure values during extraction:**

   The results showed that vowels and speech style significantly influenced spectral measure values, with results depending on the tested measure.

2. **Effectiveness of spectral measures in speaker discrimination (individually and in combinations):**

   Spectral measures, especially those incorporating formant centre frequencies, increased the classification rates for deafferenting individual speakers (see Table 7.1 and 7.9).

   The study found that formant amplitude and spectral peak measures outperformed LPC-derived and spectrally derived bandwidth measures.

   The combination of spectral measures considerably enhanced model performance, with models including eight spectral measures performing up to 12 times better than chance (for all uncorrelated measures).

3. **Factors affecting discriminant values of spectral measures:**

   When delving deeper into vowels, /u:/, /e/, /a:/, and /ɪ/ yield the highest CR results (see table 7.8).

   Front vowels perform slightly better than back vowels.

   Speech styles with the highest CRs were wordlists.

The study showed that spectral measure analysis performed equally well across all three varieties, indicating no significant advantage for any particular variety.

In conclusion, the study demonstrates that spectral measures, when combined together with other measures or centre formant frequencies, effectively differentiate individual speakers in acoustic analysis, influenced by factors such as vowel choice and data elicitation mode/speech style (based on LDA classification). The study provides insights into spectral measures' practical application in speaker differentiation and speaker recognition, with the potential for further research to enhance accuracy.

.

9.1.3 Article 3

Article 3 examines speaker discriminatory power achieved by combining spectral features from Articles 1 and 2, assessing if this combination enhances system performance.

1. **Feature combinations with high discriminatory power:**

   Formant amplitudes were found to be the most accurate features, surpassing the best-performing spectral moment, centre of gravity.

   Combining within-formant spectral moments with spectral measures significantly improved speaker CR's values.

2. **Optimal feature combination:**

   The optimal feature combination, consisting of the best-performing measures from both studies, achieved a CR 15.4 times above the chance level.

   The study found that adding 6 to 7 features to a model yielded the most effective results, with returns diminishing beyond this point.

3. **Factors affecting the spectral feature analysis:**

   The study found that the vowel /a:/ was the most effective in discriminating, however, when analysed as a subset, close vowels demonstrated superior discriminatory power.

   Wordlist consistently demonstrated superior performance compared to the other two methods, as demonstrated in Articles 1 and 2.

Brahmin outperformed the other two varieties, but the difference in correct CRs between the three varieties was minimal. The spectral feature analysis demonstrated consistent speaker discriminatory power across various language varieties.

In conclusion, the study highlights the effectiveness of combining spectral moments and measures with centre formant frequencies in a single system for speaker discrimination. Further research is needed to expand the system's accuracy to other languages and consider acoustic features and age/gender differences.

9.2 General Conclusion of the Study

The study embarked on addressing three rationales for favouring a human-assisted acoustic analysis-based system over an automated counterpart, as discussed in Chapter 2. The first rationale centred on the limited research into within-formant features, prompting a comprehensive examination of these features as a possibility for enhancing the effectiveness of a human-assisted system. Articles 1-3 investigated these features, both individually and in combinations while considering three factors. The research findings underscored that their inclusion, in conjunction with centre formant frequencies, led to a significant improvement in correct classification rates.

Prior literature had posited that higher formant frequencies, notably F3, would carry more speaker-specific characteristics compared to lower formants (McDougall, 2004). However, the results presented here challenged this notion. When analysing within-formant features, it became evident F1, surprisingly, harbours more speaker-specific information than the other formants. Specifically, both spectral moments and spectral measures highlighted F1's centre of gravity and formant amplitudes as the best-performing measures. This trend gained further prominence when the features were assessed for specific vowels i.e., [iː], [ɪ], [e], [ə], [aː], [o], [uː], [ʊ]. Consequently, the result indicates that, while in isolation, F3's centre formant value may exhibit a higher degree of speaker discrimination, the within-formant features offer a greater inter-speaker discrimination potential for F1.

This discrepancy might be attributed to the fact that both spectral moments and measures rely on energy concentrations surrounding formant peaks. As higher formants tend to experience more energy losses (for females), valuable information may be lost at these frequencies. This finding carries particular significance for forensic speaker comparison studies, given that most

telephonic recordings are frequency band-limited, resulting in clearer lower formant peaks and energy around these peaks than their higher counterparts.

Another compelling explanation for the divergent findings in the present study compared to others lies in the insights provided by Baumann and Belin (2010). According to their research, when distinguishing between female voices, F1 carries more distinctive information than the higher formants, a pattern not observed in male voices. This discrepancy is attributed to the fact that the higher formants of female speakers tend to carry less acoustic energy compared to their male counterparts. Consequently, while F3 and F4 may serve as effective discriminators in speaker classification for male speaker comparison studies, the present study's results suggest that the energy content in the higher formants regions may be too attenuated for practical utilization in within-formant feature-based analysis, particularly for female speakers.

A second pivotal discovery in the study highlighted the substantial advantage of close vowels over open vowels in both spectral moments and spectral measures analysis. This phenomenon can be ascribed to several underlying factors.

To begin with, close vowels are characterized by a more constricted oral cavity, a reduced resonating space, and lower variability in their first formant (F1) (Mitsuya et al., 2015). Consequently, this might result in a more distinct distribution of energy and greater variation in the vocal tract length and positioning among different speakers (F2 and F3 variability). In contrast, open vowels exhibit higher production variability, resulting in a more open and less constrained oral cavity (Beckman et al., 1995). As a result, the reduced variability in formant frequencies for open vowels makes them less distinctive in terms of acoustic patterns. The close vowels in Marwari language occupied almost same space (see Chapter 3 for further details), which could have led speakers to have a better precision for articulation in order to differentiate phoneme representations, resulting in lower space for inter-speaker variability.

By amalgamating the main findings presented herein, an argument emerges i.e., within-formant features extracted from F1 yield substantially higher classification rates than those obtained from any other formants under investigation for female speakers. In essence, these findings collectively suggest that the efficacy of within-formant features, particularly in the context of F1 and close vowels, can be associated with the advantageous characteristics of energy distribution within lower formant regions. Notably, F1 served as a more discriminative measure for identifying female speakers in this study, aligning with Baumann and Belin's (2010)

findings that the first formant (F1) tends to provide more speaker-distinctive information for females compared to males. While the current findings lend support to potential sex differences in vowel-based cues to speaker identity, the effects of sex could not be directly ascertained due to the sole focus on female speakers. Further research through direct comparisons of open versus close vowel articulation for speaker recognition in both males and females within the same methodological paradigm is warranted to delineate a more nuanced perspective on the role of vowel quality in indexing speaker identity across the sexes.

Although the Marwari language varieties showed minimal differences in formant values, greater distinctions emerged in the spectral measures derived from these formants across varieties. This suggests the spectral characteristics encode more variety-specific information compared to overall formant positioning. Nevertheless, the classifier performance remained consistent regardless of the language variety, indicating speaker discriminability was preserved even for varieties with sub phonemic distinctions. While spectral measures revealed cross-varietal nuances, the speaker classification rates implied that higher-level indexical cues transcend fine-grained spectral variability across closely related varieties. This lends promise to the findings and proposed methods being generalisable beyond Marwari to other languages.

The resilience of the method to within-language dialectal variation indicates the approach could be applicable to multiple languages, rather than being restricted to the Marwari data under consideration. Based on the outcomes of this study, it is possible to speculate that in the future, one variety of Marwari may be used to train any model, and speakers from another variety may be assessed using this trained model.

This underscores the first as well as the second motivation for this study: the assessment of the role played by within-formant features in the realm of speaker comparison and utilizing more interpretable spectral features to improve the transparency and reliability in contrast to the ASR systems. Through a meticulous examination of these features, the study endeavours to augment the efficiency of AcPA-based speaker comparison systems, concurrently emphasizing the importance of within-formant features. The study's contribution extends into the domain of forensic research, with a particular focus on female speakers. Moreover, the study offers valuable insights into which features are more likely to yield superior results, potentially making it interpretable, reliable and replicable when applied to other languages.

9.3 Limitations

In summarizing the limitations across the articles, several common constraints emerge. First, the complexity of the data extraction process, which relied on various settings, introduces a vulnerability to potential alterations in the results. Second, the arduous nature of gathering fieldwork data in rural areas, coupled with unanticipated issues related to speaker accommodations during fieldwork, raised concerns about the integrity of wordlist data. Additionally, the Marwari language's lower frequency of specific vowels, namely /ʊ/ and /u:/ translated to fewer instances per speaker, impacting the vectors utilised in the linear discriminant analysis. Furthermore, the non-laboratory settings in which data collection occurred may have influenced the acoustic representation of features and might have introduced additional variability into the extraction process. There is a clear need to create and test subsets of participants based on variables such as age, gender and education to validate – or otherwise - the efficacy of the identified features. The original research plan, altered due to the COVID-19 pandemic, underscores the importance of considering different time periods in future studies. It is advisable to enrich the analytical system by incorporating supplementary features such as voice quality, phonation types, articulation rate, fundamental frequencies and long-term formant distributions, among others, to enhance the comprehensiveness of the research.

One notable limitation of this analysis for forensic speaker comparison is that it performs least well on conversational speech which is the most comparable to genuine forensic samples. In contrast, it functions optimally on wordlists, the least likely input in casework. However, for applications such as voice-authenticated banking, and other customer services, wordlists may be the closest equivalent to the typical input.

Overall, while this method shows promise with some speech types, more research is needed to determine if performance could be improved for natural conversation before it could be reliably applied to forensic comparisons.

9.4 Implications

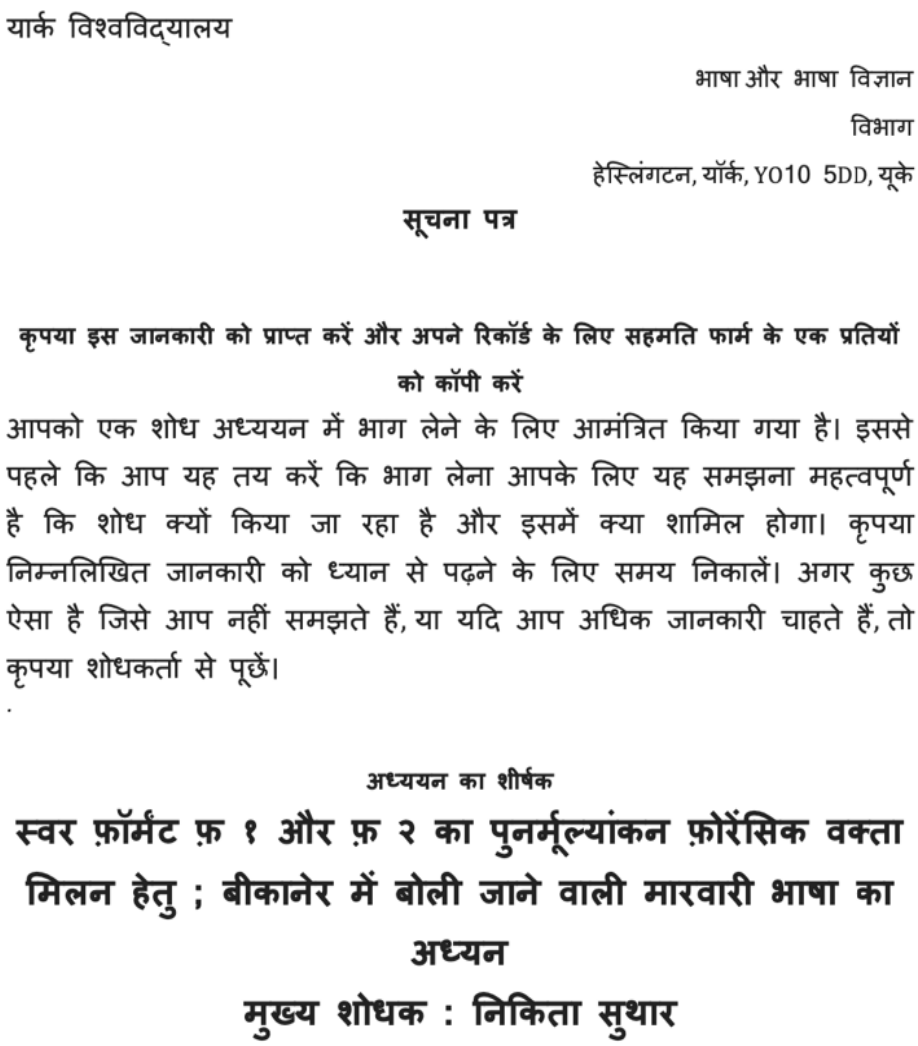
The aim of the study is to stimulate dialogue and address the challenges inherent in human-assisted speaker comparison work through its findings. Based on the research outcomes, a strong recommendation is made to incorporate and assess the integration of within-formant

metrics into human-assisted acoustic phonetic-based speaker comparison investigations, by adding them to the existing techniques (but not as a standalone system).

Additionally, this work sheds light on the outcomes obtained from female participants of a lesser-known language, demonstrating differences compared to mainstream languages such as that spoken by native English males.

10 Appendix

10.1 Additional images and tables from Chapters 4 and 5



यार्क विश्वविद्यालय

भाषा और भाषा विज्ञान
विभाग
हेस्लिंगटन, यॉर्क, YO10 5DD, यूके

सूचना पत्र

कृपया इस जानकारी को प्राप्त करें और अपने रिकॉर्ड के लिए सहमति फार्म के एक प्रतियों को कॉपी करें

आपको एक शोध अध्ययन में भाग लेने के लिए आमंत्रित किया गया है। इससे पहले कि आप यह तय करें कि भाग लेना आपके लिए यह समझना महत्वपूर्ण है कि शोध क्यों किया जा रहा है और इसमें क्या शामिल होगा। कृपया निम्नलिखित जानकारी को ध्यान से पढ़ने के लिए समय निकालें। अगर कुछ ऐसा है जिसे आप नहीं समझते हैं, या यदि आप अधिक जानकारी चाहते हैं, तो कृपया शोधकर्ता से पूछें।

अध्ययन का शीर्षक

स्वर फ़ॉर्मंट फ़ १ और फ़ २ का पुनर्मूल्यांकन फ़ोरेंसिक वक्ता मिलन हेतु ; बीकानेर में बोली जाने वाली मारवारी भाषा का अध्यन

मुख्य शोधक : निकिता सुथार

शोध किस बारे में है?

इस शोध का उद्देश्य विभिन्न विशेषताओ में से एक विशेषता की जांच करना है जो कई संदिग्ध भाषण नमूनों से किसी व्यक्ति की आवाज की पहचान करने में सहायक हो सकती

Figure 10.1 An example of the consent form provided to the participant in Devanagari and Roman script.

Table 10.1 Wordlist used for the study. The wordlist follows CVC rules where both consonants preceding and following the vowels are obstruents.

Number	/e/	/ɑ:/	/o/	/i:/	/ɪ/	/u:/	/ʊ/	/ə/
1	ɖeɖ	dʰɑ:p	gʰoʈo	ɖi:ɖ	ɖʰɪkɳo	ɖu:dʒ	bʊɖɖɪ	səsʊrɑ:l
2	bʰeɖ	bɑ:bo	dʒʰoʈo	bi:blɪ	ʃɪkər	ɖʰu:ɖ	gʰʊmər	səɖək
3	ɖeɖɪ	ɖɑ:ɖo	tʃʰoʈo	bʰebʰi:ʈ	ʃɪʃʊ	ɖʰu:ɖʰ	ʃʊkər	nukəs
4	keɖ	sɑ:s	boɖo	ʃi:ʃo	sɪski:	bu:dʒʰəɳo	dʒʰʊkəɳo	rəɖək
5	bʰebʰi:t	tʃʰɑ:tʃʰ	bobo	si:ʃʊ	kɪʈʰ	su:g	dʒʊkʰɑ:m	rədʒək
6	veɖ	sɑ:g	voɖo	ɖi:se	pɪtʃəkko	dʒʰu:ʈʰ	kʊɖəɳo	pəg
7	ʈʰeʈʰ	bɑ:dʒ	koɖ	dʒi:dʒo	dʒɪdʒi:	ʈʰu:k	ɖʊgər	dʒək
8	fefɪ	ɖɑ:g	sotʃ	ʈi:p	ʈɪrəʈʰ	dʒu:dʒʰ	sʊkʰ	dʒedʒəs
9	dʒedʒʰ	kʰɑ:kʰ	dʒovən	pi:ɖ	gʰɪsəɳo	ɖu:dʒ	pʊgəl	dʒɪdʒək
10	kesʰ	pɑ:k	ʈoɖ	bi:dʒ	gɪgləjɑ:	ɖʰu:p	gʊsso	gʰəgʰərɪ

Table 10.2 p-values for three different models for spectral moments

Feature	Model	P-value		
		Wordlist	Story	Conversation
F1- m_1	Variety	<< .0001	<< .0001	0.0006
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.0097	0.0689
F2- m_1	Variety	<< .0001	<< .0001	<< .0001
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.0001	0.0002
F3- m_1	Variety	<< .0001	<< .0001	<< .0001
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	<< .0001	0.3219
F4- m_1	Variety	<< .0001	<< .0001	<< .0001
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.4465	0.2530	0.3718
F1- m_2	Variety	<< .0001	<< .0001	0.0032
	Vowel	0.0162	0.0116	0.0218
	Variety *Vowel	0.0003	0.0411	0.1012
F2- m_2	Variety	0.0062	0.0188	0.0001
	Vowel	0.0002	<< .0001	0.0083
	Variety *Vowel	0.2292	0.0925	0.1052
F3- m_2	Variety	0.0083	0.0219	0.1193
	Vowel	<< .0001	0.0311	<< .0001
	Variety *Vowel	0.0788	0.2092	0.3039
F4- m_2	Variety	0.0490	0.1250	0.7517
	Vowel	0.0007	0.0012	0.3024
	Variety *Vowel	0.0793	0.5968	0.5033
F1- m_3	Variety	0.1207	0.7766	0.5842
	Vowel	0.0001	0.1103	0.9447
	Variety *Vowel	0.0041	0.6868	0.5484
F2- m_3	Variety	0.0808	0.5492	0.2249
	Vowel	0.0001	0.0933	0.0075
	Variety *Vowel	0.1987	0.1265	0.3971
F3- m_3	Variety	0.0020	0.4072	0.8528
	Vowel	<< .0001	0.1767	0.0151
	Variety *Vowel	0.5675	0.3737	0.2207
F4- m_3	Variety	0.0703	0.9750	0.3860
	Vowel	0.0504	0.7573	0.3571
	Variety *Vowel	0.4749	0.0680	0.2637
F1- m_4	Variety	<< .0001	<< .0001	0.6458
	Vowel	0.0010	0.1371	0.5419
	Variety *Vowel	0.3812	0.1732	0.8478
F2- m_4	Variety	0.4370	0.3934	0.7509
	Vowel	<< .0001	0.0586	<< .0001
	Variety *Vowel	0.7775	0.1340	0.0821
F3- m_4	Variety	0.6642	0.6947	0.8754
	Vowel	<< .0001	0.0007	0.0658
	Variety *Vowel	0.1430	0.3396	0.6188
F4- m_4	Variety	0.1373	0.6132	0.2322
	Vowel	0.0025	0.2393	0.3836
	Variety *Vowel	0.0405	0.3251	0.4868

Table 10.3 p-values for three different models for spectral measures

Feature	Model	P-value		
		Wordlist	Story	Conversation
F1-A1	Variety	<< .0001	<< .0001	0.4377
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.1422	0.0038	0.8749
F2-A2	Variety	<< .0001	<< .0001	0.4576
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.4479	0.0021	0.8642
F3-A3	Variety	<< .0001	<< .0001	0.5980
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	<< .0001	0.9737
F4-A4	Variety	0.3791	<< .0001	0.5420
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.0935	0.0003	0.9986
F1-SB1	Variety	<< .0001	<< .0001	<< .0001
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.2087	0.7343
F2-SB2	Variety	0.8642	0.0127	<< .0001
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.0022	0.0026	0.2639
F3-SB3	Variety	0.7855	0.1816	<< .0001
	Vowel	<< .0001	0.0033	0.0016
	Variety *Vowel	<< .0001	0.0011	0.6358
F4-SB4	Variety	0.0031	0.1161	<< .0001
	Vowel	<< .0001	0.0001	0.0298
	Variety *Vowel	0.0035	0.5540	0.7911
F1-LB1	Variety	0.4343	<< .0001	0.2611
	Vowel	<< .0001	<< .0001	0.0030
	Variety *Vowel	0.0872	0.0017	0.9958
F2-LB2	Variety	0.0130	0.2906	0.1540
	Vowel	<< .0001	0.0654	0.0018
	Variety *Vowel	0.0042	0.0129	0.9554
F3-LB3	Variety	0.0038	0.0671	0.4090
	Vowel	0.0469	<< .0001	0.0002
	Variety *Vowel	0.0258	0.2259	0.9867
F4-LB4	Variety	0.0005	<< .0001	0.3361
	Vowel	<< .0001	0.0004	0.0791
	Variety *Vowel	0.1692	0.0644	0.9888
F1-SP1	Variety	<< .0001	<< .0001	0.9025
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.1727	0.9996
F2-SP2	Variety	<< .0001	<< .0001	0.8865
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.0001	0.9996
F3-SP3	Variety	<< .0001	<< .0001	0.9878
	Vowel	<< .0001	<< .0001	0.1200
	Variety *Vowel	0.0004	<< .0001	0.9969
F4-SP4	Variety	<< .0001	<< .0001	0.5876
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.3017	0.1222	0.9998
A2-A1	Variety	<< .0001	<< .0001	0.7538
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.0001	0.0030	0.9945

A3- A1	Variety	0.0057	<< .0001	0.9651
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	<< .0001	0.9767
A4- A1	Variety	<< .0001	0.0158	0.8890
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	<< .0001	0.0139	0.9884
A3-A2	Variety	<< .0001	0.2016	0.7655
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.0233	0.0036	0.9680
A4- A2	Variety	<< .0001	<< .0001	0.8808
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.2200	0.1040	0.9564
A4- A3	Variety	<< .0001	<< .0001	0.9367
	Vowel	<< .0001	<< .0001	<< .0001
	Variety *Vowel	0.0053	<< .0001	0.9991

Table 10.4 Best performing spectral features between amplitude, bandwidth and spectral peaks for each vowel category for every variety (Article 2, section 7.8.7)

Variety	**Brahmin**				**Jaat**				**Bishnoi**			
Type	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean	Wordlist	Story	Conversation	Mean
Mean	1.8	2.1	1.5		2.1	1.9	1.4		2.0	1.9	1.9	
/aː/	2.5 LB1	2.5 SP3	2.4 A2 A4	2.5	3.4 SP1 SP4	3.4 SP2	2.4 A3	3.1	2.7 SP4	2.6 A1 A2A1		2.6
/e/	3.4 SP1 SP2	2.9 A3A1	2.8 A2	3	3.1 A3A1	3.1 SP4	2.1 SP2	2.8	7.6 SB3	3.6 SP3		5.6
/ə/	2 A4	4.1 SP1	2.6 A2	3		4 A4	1.8 A2A1	2.9	3.5 A3 A4	2.5 A2		2.9
/iː/	3.2 A1	3.8 A1 A3A2 A3A2	2.2 SB1	3.1		2.4 A4A3	2.6 SP4	2.5		2.7 A2	2.9 SP1	2.8
/ɪ/	3.4 SB1	3.3 A2A1	1.8 A2 LB3 A4A1	2.9	2.8 A2 A3 SP1 SP3	3.37 A1	2.2 SP1 A4A3	2.8	2.9 A1 A2 SB2	2.9 A2A1		2.9
/o/	2.7 SP3	3.7 A1	2.2 A3A1	2.9		2.7 SB2	1.6 A2 SB1 SB4	2.2		2.4 A4A1	3.6 A4	3
/uː/	3.2 A1	2.5 A3A2	1.7 A4 SB1 A4A1	2.4	2.5 A4	2.5 A3	2.2 SP1	2.4	3.9 SB1	3.4 LB2		3.6
/ʊ/		3 A2A1	2 A1 A2 SB3	2.5	3.1 A1	3.4 A1 A2	1.7 A3 A4 SP3 SB3 A3A1	2.7	2.9 SB2	2.2 A4		2.6

10.2 Some additional graphs of the performance of individual spectral features for different vowels from Article 3

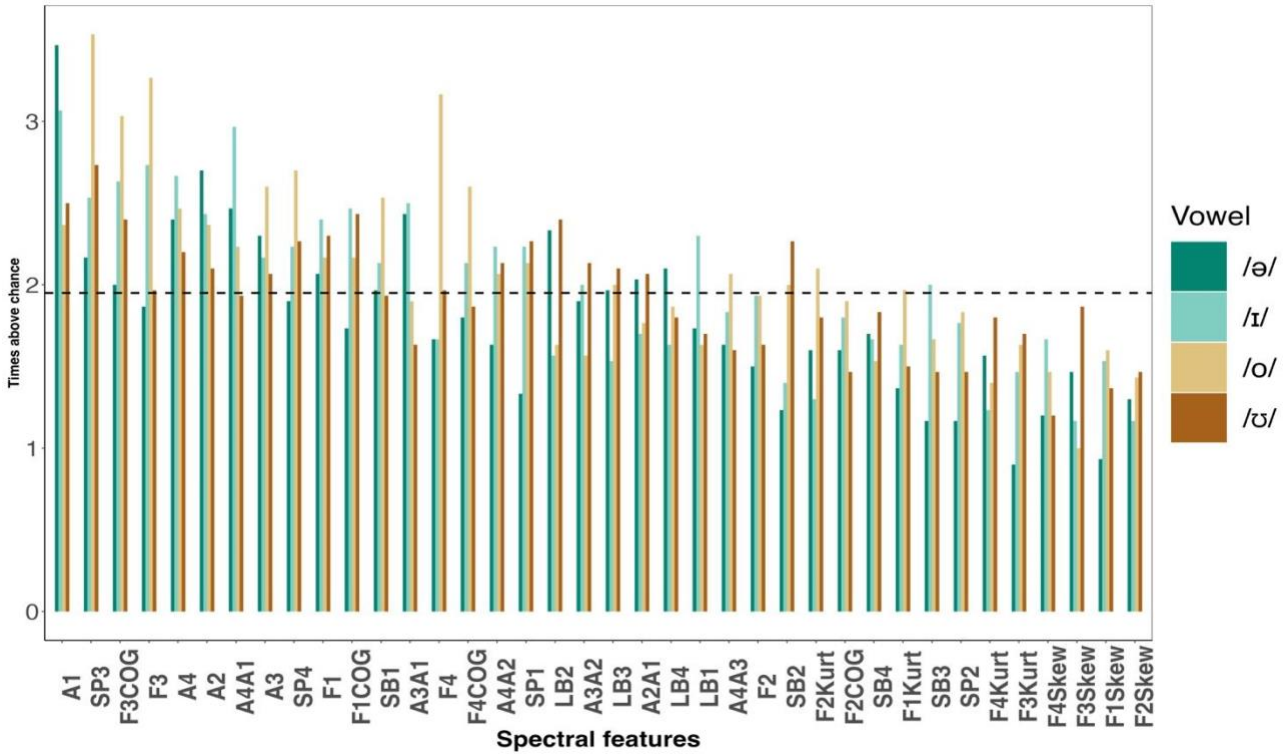


Figure 10.2 Individual performances of vowels /a:/, /e/, /i:/, /u:/ of for story data.

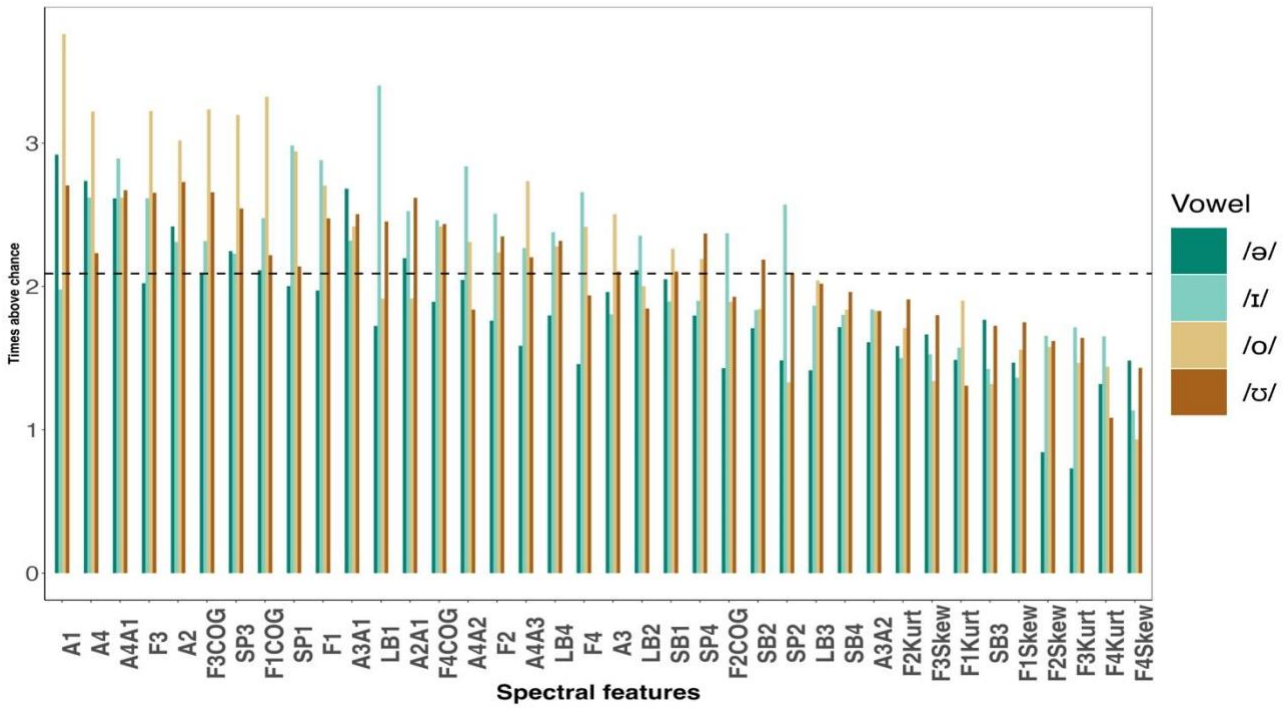


Figure 10.3 Individual performances of vowels /o/, /ʊ/, /ɪ/, /ə/ of for story data.

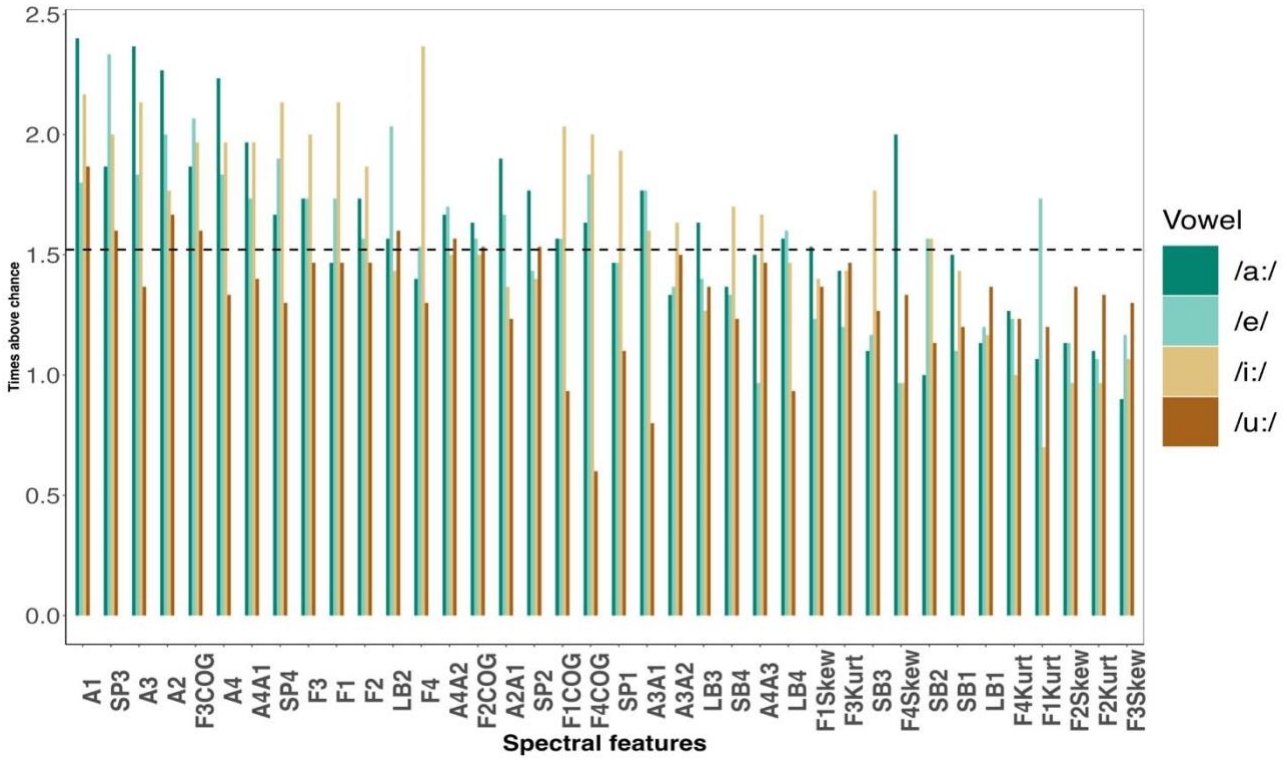


Figure 10.4 Individual performances of vowels /a:/, /e/, /i:/, /u:/ of for Conversation data.

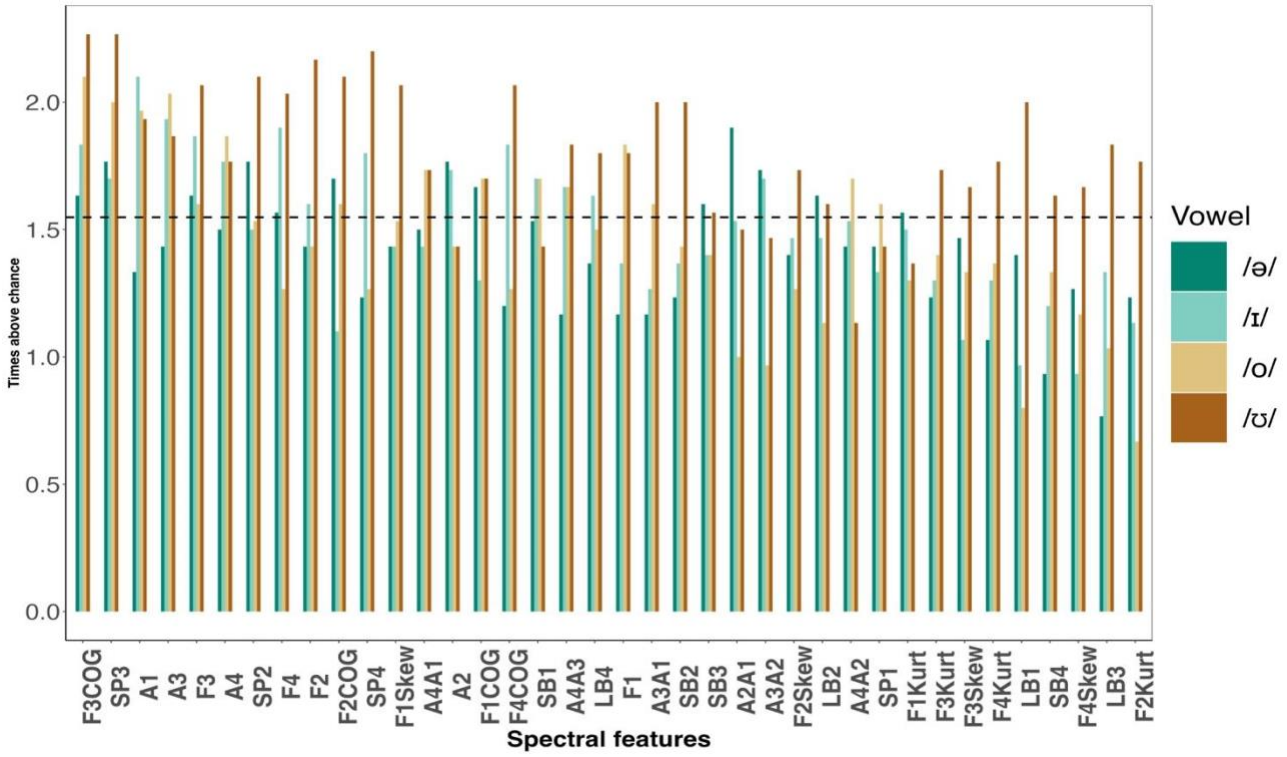


Figure 10.5 Individual performances of vowels /o/, /ʊ/, /ɪ/, /ə/ of for Conversation data.

11 List of Abbreviations

Abbreviation	Full Form	Abbreviation	Full Form
1 (P)	1st Person	F1Kurt	Kurtosis for first formant
2 (P)	2nd Person	F1SD	Standard deviation for first formant
3 (P)	3rd Person	F1Skew	Skewness for first formant
A1	Formant-amplitude of the first formant	F2	Second formant
A2	Formant-amplitude of the second formant	F2COG	Centre of gravity for second formant
A3	Formant-amplitude of the third formant	F2Kurt	Kurtosis for second formant
A4	Formant-amplitude of the fourth formant	F2SD	Standard deviation for second formant
AcPA	Acoustic Phonetic Analysis Only	F2Skew	Skewness for second formant
Amp	Amplitude	F3	Third formant
ANOVA	Analysis of variance	F3COG	Centre of gravity for third formant
ASR	Automatic Speaker Recognition	F3Kurt	Kurtosis for third formant
AuPA	Auditory Phonetic Analysis Only	F3SD	Standard deviation for third formant
AuPA+AcPA	Auditory Phonetic cum Acoustic Phonetic Analysis	F3Skew	Skewness for third formant
Aux.	Auxiliary	F4	Fourth formant
COG	Centre of gravity (first spectral moment)	F4COG	Centre of gravity for fourth formant
CR	Classification rate	F4Kurt	Kurtosis for fourth formant
CVC	Consonant-vowel-consonant	m_2	Second spectral moment (standard deviation)
dB	Decibel	m_3	Third spectral moment (skewness)
dBFS	Decibel full scale	m_4	Fourth spectral moment (kurtosis)
DS	Different speaker	MFCCs	Mel Frequency Cepstral Coefficients
F4SD	Standard deviation for fourth formant	Neg.	Negative
F4Skew	Skewness for fourth formant	P.	Plural
F5	Fifth formant	Past	Past Tense

Fut.	Future Tense
GMM-UBM	Gaussian mixture model - Universal background model
HASR	Automatic Speaker Recognition System with Human Assistance
HMM	Hidden Markov Model
Hon.	Honorific Marker
Hz	Hertz
i-vector	Identity-vector
Kurt	Kurtosis (fourth spectral moment)
LB	LPC-bandwidth
LB1	LPC-bandwidth of first formant
LB2	LPC-bandwidth of second formant
LB3	LPC-bandwidth of third formant
LB4	LPC-bandwidth of fourth formant
LDA	Linear discriminant analysis
LLR	Log-likelihood ratio
lmer	Linear mixed-effect regression
LPC	Linear predictive coding
LR	Likelihood ratio
M.	Masculine
m_1	First spectral moment (centre of gravity)
F.	Feminine
f0	Fundamental frequency
F1	First formant
F1COG	Centre of gravity for first formant
PCA	Principal component analysis
PLDA	Probabilistic linear discriminant analysis
Prep.	Preposition
Pres.	Present Tense
S.	Singular
SB	Spectral bandwidth
SB1	Spectral bandwidth of first formant
SB2	Spectral bandwidth of second formant
SB3	Spectral bandwidth of third formant
SB4	Spectral bandwidth of forth formant
SD	Standard deviation (second spectral moment)
SFA	Spectral feature analysis
Skew	Skewness (third spectral moment)
SM	Spectral moments
SMA	Spectral moment analysis
SOV	Subject-Object-Verb
SP	Spectral peak
SP1	Spectral peak of first formant
SP2	Spectral peak of second formant
SP3	Spectral peak of third formant
SP4	Spectral peak of fourth formant
SS	Same speaker

References

Aaltonen, O. (1985). The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. *Journal of Phonetics*, *13*(1), 1-9. https://doi.org/10.1016/S0095-4470(19)30721-1

Abbi, A. (2010). Vanishing diversities and submerging identities an Indian case. In A. Sarangi (Ed.), *Language and politics in India* (pp. 299-311). Oxford University Press. http://www.linguapax.org/wp-content/uploads/2015/03/2_abbi.pdf

Adi-Bensaid, L., & Tobin, Y. (2010). Is there compensatory vowel lengthening in the language acquisition of a child with a cochlear implant? *Poznan Studies in Contemporary Linguistics*, *46*(3), 255-274. https://doi.org/10.2478/v10010-010-0015-5

Ainsworth, W. A., & Millar, J. B. (1972). The effect of relative formant amplitude on the perceived identity of synthetic vowels. *Language and Speech*, *15*(4), 328-341. https://doi.org/10.1177/002383097201500403

Aithal, P. S., & Aithal, S. (2020). Analysis of the Indian National Education Policy 2020 towards achieving its objectives. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *5*(2), 19-41. https://doi.org/10.2139/ssrn.3676074

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *53*(1), 109-122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists* (2nd ed.). John Wiley & Sons, Ltd.

Alam, M. J., Kenny, P., & Stafylakis, T. (2015). Combining amplitude and phase-based features for speaker verification with short duration utterances. *Proceedings of Interspeech 2015* (pp. 249-253). Dresden, Germany. https://doi.org/10.21437/interspeech.2015-94

Alekseevna, R. E., & Sergeevna, K. A. (2021). Education system against language transmission? Case of minority languages in India. *Tomsk Journal of Linguistics and Anthropology*, *33*(3), 70-80. https://doi.org/10.23951/2307-6119-2021-3-70-80

Alison, L., Sarangi, S., & Wright, A. (2008). Human rights is not enough: The need for demonstrating efficacy of an ethical approach to interviewing in India. *Legal and Criminological Psychology*, *13*(1), 89-106. https://doi.org/10.1348/135532506X157737

Alwan, A., Narayanan, S., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *The Journal of the Acoustical Society of America*, *101*(2), 1078-1089. https://doi.org/10.1121/1.417972

Assmann, P. F. (1991). The perception of back vowels: centre of gravity hypothesis. *The Quarterly Journal of Experimental Psychology Section A*, *43*(3), 423-448. https://doi.org/10.1080/14640749108400980

Atal, B. S., & Schroeder, M. (1978). Linear prediction analysis of speech based on a pole-zero representation. *The Journal of the Acoustical Society of America*, *64*(5), 1310-1318. https://doi.org/10.1121/1.382117

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048-3058. https://doi.org/10.1121/1.2188331

Babel, M. E. (2009). *Phonetic and social selectivity in speech accommodation*. [PhD Thesis, University of California] Berkeley, USA. http://escholarship.org/uc/item/1mb4n1mv

Babel, M. E. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177-189. https://doi.org/10.1016/j.wocn.2011.09.001

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The*

Journal of the Acoustical Society of America, *106*(2), 1054-1063. https://doi.org/10.1121/1.427115

Barber, C., Beal, J. C., & Shaw, P. A. (2012). *The English language: A historical introduction* (5th ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511817601

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1 - 48. https://doi.org/10.18637/jss.v067.i01

Batra, P. (2020). Echoes of 'coloniality' in the episteme of Indian educational reforms. *On Education: Journal for Research and Debate*, *3*(7), 1-9. https://doi.org/10.17899/on_ed.2020.7.3

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110-120. https://doi.org/10.1007/s00426-008-0185-z

Becker, T., Jessen, M., & Grigoras, C. (2008, September 22-26, 2008). Forensic speaker verification using formant features and Gaussian mixture models. *Proceedings of Interspeech 2008, 9th Annual Conference of the International Speech Communication Association* (pp.1505-1508), Brisbane, Australia. https://doi.org/10.21437/Interspeech.2008-432

Beckman, M. E., Jung, T. P., Lee, S. h., Jong, K. d., Krishnamurthy, A. K., Ahalt, S. C., Cohen, K. B., & Collins, M. J. (1995). Variability in the production of quantal vowels revisited. *The Journal of the Acoustical Society of America*, *97*(1), 471–490. https://doi.org/10.1121/1.412945

Belsley, D.A. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. David A. Belsley, Edwin Kuh, Roy E. Welsch (Eds.). Wiley-Interscience. https://ebookcentral.proquest.com/lib/york-ebooks/reader.action?docID=226535

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145-204. https://doi.org/10.1017/S004740450001037X

Beshears, A. (2017). *The demonstrative nature of the Hindi/Marwari correlative* [PhD Thesis, Queen Mary University of London]. http://qmro.qmul.ac.uk/xmlui/handle/123456789/30629

Béteille, A. (1967). Race and descent as social categories in India. *Daedalus*, *96*(2), 444-463. https://www.jstor.org/stable/20027046

Bhatt, S., Jain, A., & Dev, A. (2021). Feature extraction techniques with analysis of confusing words for speech recognition in the Hindi language. *Wireless Personal Communications*, *118*, 3303-3333. https://doi.org/10.1007/s11277-021-08181-0

Bhattacharya, U. (2017). Colonization and English ideologies in India: A language policy perspective. *Language Policy*, *16*, 1-21. https://doi.org/10.1007/s10993-015-9399-2

Biemans, M. A. J. (2000). *Gender variation in voice quality* [PhD Thesis, Katholieke Universiteit Nijmegen]. Nijmegen, The Netherlands (Utrecht). https://www.lotpublications.nl/Documents/038_fulltext.pdf

Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., & Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing 2004*, *2004*, 430-451. https://doi.org/10.1155/S1110865704310024

Blacklock, O. S. (2004). *Characteristics of variation in production of normal and disordered fricatives, using reduced-variance spectral methods* [PhD Thesis, University of Southampton]. http://eprints.soton.ac.uk/id/eprint/420069

Bladon, R. A. W., & Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *The Journal of the Acoustical Society of America*, *69*(5), 1414-1422. https://doi.org/10.1121/1.385824

Boedeker, P., & Kearns, N. T. (2019). Linear discriminant analysis for prediction of group membership: A user-friendly primer. *Advances in Methods and Practices in Psychological Science, 2*(3), 250-263. https://doi.org/10.1177/2515245919849378

Boersma, P., & Weenink, D. (2001). *Praat, a system for doing phonetics by computer* (6.2.23) [Computer Software]. Glot International. http://www.praat.org/

Bortlík, M. J. F. (2021). *Czech accent in English: Linguistics and biometric speech technologies* [PhD Thesis, Palacký University Olomouc]. https://theses.cz/id/s2en4i/Bortlik_2021_phd_dissertation.pdf

Bourlière, F. (1970). The assessment of biological age in man. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/37808/WHO_PHP_37.pdf?sequence=1&isAllowed=y

Braun, B., Dehé, N., Einfeldt, M., Wochner, D., & Zahner-Ritter, K. (2021, 30 August – 3 September). Testing acoustic voice quality classification across languages and speech styles. *Proceedings of Interspeech 2021* (pp. 3920-3924), Brno, Czechia. https://doi.org/10.21437/Interspeech.2021-315

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language, 20*(2-3), 230-275. https://doi.org/10.1016/j.csl.2005.08.001

Burghart, R. (1978). Hierarchical models of the Hindu social system. *Man, 13*(4), 519-536. https://doi.org/10.2307/2801246

Burris, C., Vorperian, H. K., Fourakis, M., Kent, R. D., & Bolta, D. M. (2014). Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements. *Journal of Speech, Language, and Hearing Research, 57*(1), 26-45. https://doi.org/10.1044/1092-4388(2013/12-0103)

Byrne, C., & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *International Journal of Speech, Language and the Law*, *11*(1), 83-102. https://www.equinoxpub.com/journals/index.php/IJSLL/article/view/540

Campbell Jr., J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*. https://doi.org/10.1109/5.628714

Cao, H., & Dellwo, V. (2019). The role of the first five formants in three vowels of Mandarin for forensic voice analysis. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 617-621), Melbourne, Australia. https://doi.org/10.5167/uzh-177494

Cao, W. (2018). *Short-term accommodation of Hong Kong English speakers towards native English accents and the effect of language attitudes* [PhD thesis, University of York]. White Rose eTheses Online. York. https://etheses.whiterose.ac.uk/23588/

Cardona, G., & Suthar, B. (2014). Gujarathi. In G. Cardona & D. Jain, *The Indo-Aryan languages* (pp. 722-765). Routledge.

Carlson, R., Granström, B., & Klatt, D. (1979). Vowel perception: The relative perceptual salience of selected acoustic manipulations. *Speech Transmission Laboratories (Stockholm) Quarterly Progress Report Status Report*, *20*(3-4), 73-83. http://www.speech.kth.se/qpsr

Carlson, R., Granström, G. G., & Fant, B. (1970). Some studies concerning perception of isolated vowels. *Speech Transmission Laboratory Quarterly Progress and Status Report 11*(2-3), 019-035. http://www.speech.kth.se/qpsr

Casad, E. H. (1987). *Dialect intelligibility testing* (Vol. 38). Benjamin F. Elson. https://www.sil.org/system/files/reapdata/85/37/26/85372638740600013916377001436207275325/10386.pdf

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2021). Acoustic analysis of vowel formant frequencies in genetically- related and non-genetically related speakers with implications

for forensic speaker comparison. *PLoS ONE, 16*(2), e0246645. https://doi.org/10.1371/journal.pone.0246645

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2023). On the speaker discriminatory power asymmetry regarding acoustic-phonetic parameters and the impact of speaking style. [Original Research Article]. *Frontiers in Psychology, 14*, 1101187. https://doi.org/10.3389/fpsyg.2023.1101187

Census of India. (1911). *Census Report 1911, Bikaner State*. S. G. Printing.

Chacko, S., & Ngwazah, L. (2012). *Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India, Volume 6: Marwari, Merwari, and Godwari* (J. Kelsall, Ed., Vol. 6). SIL International. https://www.sil.org/resources/publications/entry/50815

Chambers, J. K. (1994). An introduction to dialect topography. *English World-Wide: A Journal of Varieties of English*, *15*(1), 35-53. https://doi.org/10.1075/eww.15.1.03cha

Chambers, J. K. (2000). Region and language variation. *English World-Wide: A Journal of Varieties of English*, *21*(2), 169-199. https://doi.org/10.1075/eww.21.2.02cha

Chandramouli, D. C. (2011). *Census of India 2011, National Population Register & Socio-Economic and Caste Census*. Ministry of Home Affairs, Government of India. https://censusindia.gov.in/census.website/

Chelliah, S. L., & Reuse, W. J. (2011). Field preparation: Philological, practical, and psychological. In Chelliah, S. L., & Reuse, W. J. (Eds.), *Handbook of descriptive linguistic fieldwork* (pp. 93–137). Springer, Dordrecht. https://doi.org/10.1007/978-90-481-9026-3_5

Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. *The Journal of the Acoustical Society of America*, *77*(3), 789-805. https://doi.org/10.1121/1.392049

Chistovich, L. A., & Lublinskaya, V. V. (1979). The ‘center of gravity’ effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, *1*(3), 185-195. https://doi.org/10.1016/0378-5955(79)90012-1

Clermont, F., & Mokhtari, P. (1994). Frequency-band specification in cepstral distance computation. *Proceedings of the 5th Australian International Conference on Speech Science & Technology, 1* (pp. 354-359), Perth, Australia. https://www.researchgate.net/publication/271909361

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*(3), 207-238. https://doi.org/10.1177/00238309040470030101

Cohen, L., Assaleh, K., & Fineberg, A. (1992). Instantaneous bandwidth and formant bandwidth. *Proceedings of IEEE Sixth SP Workshop on Statistical Signal and Array Processing* (pp. 13-17), Victoria, BC, Canada. https://doi.org/10.1109/SSAP.1992.246837

Colton, R. H., Paseman, A., Kelley, R. T., Stepp, D., & Casper, J. K. (2011). Spectral moment analysis of unilateral vocal fold paralysis. *Journal of Voice*, *25*(3), 330-336. https://doi.org/10.1016/j.jvoice.2010.03.006

Czaplicki, B., Pape, D., Zygis, M., & Jesus, L. M. T. (2016). Acoustic evidence of new sibilants in the pronunciation of young Polish women. *Poznan Studies in Contemporary Linguistics*, *52*(1), 1-42. https://doi.org/10.1515/psicl-2016-0004

Das, R. K., & Li, H. (2020). On the importance of vocal tract constriction for speaker characterization: The whispered speech study. *Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7119-7123), Barcelona, Spain. https://doi.org/10.1109/ICASSP40776.2020.9054396

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Proceedings of IEEE Transactions on Acoustics, Speech, and Signal Processing* (pp. 357-366). https://doi.org/10.1109/TASSP.1980.1163420

Dehak, N., Dehak, R., Glass, J., Reynolds, D., & Kenny, P. (2010). Cosine similarity scoring without score normalization techniques. *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2010*) (paper 15). https://www.isca-speech.org/archive/odyssey_2010/dehak10_odyssey.html

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *Proceedings of IEEE Transactions on Audio, Speech, and Language Processing* (pp. 788-798). https://doi.org/10.1109/TASL.2010.2064307

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, an international review*, *6*(44), 29-68. https://doi.org/10.1515/ling.1968.6.44.29

Diez, D., Cetinkaya-Rundel, M., & Dorazio, L. (2015). *Advanced high school statistics Second Edition.* OpenIntro, Incorporated. http://www.openintro.org/redirect.php?go=ahss&referrer=ahss2_pdf

Dimitriadis, D., Maragos, P., & Potamianos, A. (2005). Auditory Teager energy cepstrum coefficients for robust speech recognition. *Proceedings of Interspeech 2005* (pp. 3013-3016). https://doi.org/10.21437/Interspeech.2005-142

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (19984). Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)* (paper 0608), Sydney, Australia. https://doi.org/10.21437/ICSLP.1998-244

Doddington, G. R. (1985). Speaker recognition: Identifying people by their voices. *Proceedings of IEEE*, *73*(11), 1651-1664. https://doi.org/10.1109/PROC.1985.13345

Drager, K., & Hay, J. (2012). Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change*, *24*(1), 59-78. https://doi.org/10.1017/S0954394512000014

Duckworth, M., McDougall, K., de Jong, G., & Shockey, L. (2011). Improving the consistency of formant measurement. *International Journal of Speech, Language & the Law*, *18*(1), 35-51. https://doi.org/10.1558/ijsll.v18i1.35

Earnshaw, K. (2021). Examining the implications of speech accommodation for forensic speaker comparison casework: A case study of the West Yorkshire FACE vowel. *Journal of Phonetics*, *87*, 1-15. https://doi.org/10.1016/j.wocn.2021.101062

Ekaterini, D., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology*, *25*(3), 335-354. https://doi.org/10.1044/2015_AJSLP-15-0020

Endres, W., Bambach, W., & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of the Acoustical Society of America*, *49*(6B), 1842-1848. https://doi.org/10.1121/1.1912589

Eriksson, E. J., Cepeda, L. F., Rodman, R. D., McAllister, D. F., Bitzer, D., & Arroway, P. (2004). Cross-language speaker identification using spectral moments. In P. Branderud & H. Traunmüller (Eds.), *Proceedings of the XVIIth Swedish Phonetics Conference FONETIK 2004* (pp. 76-79), Stockholm University, Sweden. https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-2282

Eriksson, E. J., Cepeda, L. F., Rodman, R. D., Sullivan, K. P. H., McAllister, D. F., Bitzer, D., & Arroway, P. (2004). Robustness of spectral moments: A study using voice imitations. *Proceedings of the Tenth Australian International Conference on Speech*

Science and Technology (pp. 259-264). http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-2283

Fant, G. (1962). Formant bandwidth data. *Speech Transmission Laboratory Quarterly Progress and Status Report 2*, *3*, 1–3. http://www.speech.kth.se/prod/publications/files/qpsr/1962/1962_3_1_001-002.pdf

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations* (Vol. 2). De Gruyter Mouton. https://doi.org/10.1515/9783110873429

Fant, G. (1972). Vocal tract wall effects, losses, and resonance bandwidths. *Speech Transmission Laboratory Quarterly progress and status report*, *13*(2-3), 028-052. https://www.ece.uvic.ca/~bctill/papers/numacoust/Fant_1972.pdf

Fant, G., & Artony, M. (1963). Formant amplitude measurements. *Speech Transmission Laboratories (Stockholm) Quarterly Progress Report*, *4*(1), 1-5. http://www.speech.kth.se/qpsr

Fant, G., Fintoft, K., Liljencrants, J., Lindblom, B., & Martony, J. (1963). Formant-amplitude measurements. *The Journal of the Acoustical Society of America*, *35*(11), 1753-1761. https://doi.org/10.1121/1.1918812

Feng, Y., Hao, G. J., Xue, S. A., & Max, L. (2011). Detecting anticipatory effects in speech articulation by means of spectral coefficient analyses. *Speech Communication*, *53*(6), 842-854. https://doi.org/10.1016/j.specom.2011.02.003

Firc, A., & Malinka, K. (2022). The dawn of a text-dependent society: Deepfakes as a threat to speech verification systems. *Proceedings of The 37th ACM/SIGAPP Symposium on Applied Computing* (pp. 1646-1655), Virtual Event. https://doi.org/10.1145/3477314.3507013

Fischer-Jørgensen, E. (1968). Phonetic analysis of breathy (murmured) vowels in Gujarati*. *Annual Report of the Institute of Phonetics University of Copenhagen*, *2*, 35-85. https://doi.org/10.7146/aripuc.v2i.130674

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics*, *8*(4), 376-386. https://doi.org/10.1111/j.1469-1809.1938.tb02189.x

Fitch, W. T. (2010). *The evolution of language*. Cambridge University Press.

Flanagan J.L. & Landgraf L.L. (1968). Self-oscillating source for vocal-tract synthesizers. *IEEE Transactions on Audio and Electroacoustics*, *16* (1), 57-64, 10.1109/TAU.1968.1161949

Fleischer, M., Pinkert, S., Mattheus, W., Mainka, A., & Mürbe, D. (2015). Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall. *Biomech Model Mechanobiol, 14*(4), 719-733. https://doi.org/10.1007/s10237-014-0632-2

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, *84*(1), 115-123. https://doi.org/10.1121/1.396977

Foulkes, P., & French, P. (2012). Forensic speaker comparison: A linguistic-acoustic perspective. In L. M. Solan & P. M. Tiersma (Eds.), *The Oxford handbook of language and law* (pp. 558-572). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0041

Fox, R. A., Jacewicz, E., & Chang, C. Y. (2011). Auditory spectral integration in the perception of static vowels [Research Article]. *Journal of Speech, Language, and Hearing Research, 54*(6), 1667-1681. https://doi.org/10.1044/1092-4388(2011/09-0279)

French, J. P., Nolan, F., Foulkes, P., Harrison, P., & McDougall, K. (2010). The UK position statement on forensic speaker comparison: A rejoinder to Rose and Morrison.

International Journal of Speech, Language and the Law, *17*(2), 143–152. https://doi.org/10.1558/ijsll.v17i1.143

French, J. P. (2017). A developmental history of forensic speaker comparison in the UK. *English Phonetics, 21*, 271-286. https://eprints.whiterose.ac.uk/117763/

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, *15*(2), 87-105. https://www.jstor.org/stable/30029508

Gish, H., & Schmidt, M. (1994). Text-independent speaker identification. *IEEE Signal Processing Magazine, 11*(4), 18-32. https://doi.org/10.1109/79.317924

Gold, E., & French, J. P. (2019). International practices in forensic speaker comparisons: Second survey. *International Journal of Speech, Language and the Law*, *26*(1), 1-20. https://doi.org/10.1558/ijsll.38028

Gold, E., & French, J. P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, *18*(2), 293-307. https://doi.org/10.1558/ijsll.v18i2.293

Gold, E., French, J. P., & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics* (pp. 1-8), Montreal, Canada. https://doi.org/10.1121/1.4800285

Gonzalez-Rodriguez, J. (2011). Speaker recognition using temporal trajectories in linguistic units: The case of formant and formant-bandwidth trajectories. In P. Cosi, R. D. Mori, G. D. Fabbrizio, & R. Pieraccini (Eds.), *Proceedings of Interspeech 2011* (pp. 133-136), Florence, Italy. https://doi.org/10.21437/Interspeech.2011-48

Google Maps. (2023). *Bikaner District.* Retrieved 24th September 2023 from https://www.google.com/maps

Gusain, L. (2004). *Marwari* (1st ed.). LINCOM.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Harrington, J., Palethorpe, S., & Watson, C. I. (2007). Age-related changes in fundamental frequency and formants: A longitudinal study of four speakers. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)* (pp. 2753-2756), Antwerp, Belgium. https://doi.org 10.21437/Interspeech.2007-716

Harrison, P. (2013). *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements* [PhD Thesis, University of York]. White Rose eTheses Online.

Harrison, P. (2019). *Forensic Plugin: Praat Script* (2.2.1) [Computer Software]. Praat.

Harrison, P. (2021). *Praat script to generate various spectral analyses around formant peaks* (1.1) [Computer Software]. Praat.

Hawks, J. W., & Miller, J. D. (1995). A formant bandwidth estimation procedure for vowel synthesis [43.72.Ja]. *The Journal of the Acoustical Society of America*, *97*(2), 1343-1344. https://doi.org/10.1121/1.412986

Hildebrandt, K. A. (2005). A phonetic analysis of Manange segmental and suprasegmental properties. *Linguistics of the Tibeto-Burman Area*, *28*(1). http://sealang.net/sala/archives/pdf4/hildebrandt2005phonetic.pdf

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, *37*(4), 769-778. https://doi.org/10.1044/jshr.3704.769

Hillenbrand, J., & Houde, R. A. (1995). Vowel recognition: Formants, spectral peaks, and spectral shape. *The Journal of the Acoustical Society of America*, *98*(5_Supplement), 2949-2949. https://doi.org/10.1121/1.414088

Hillenbrand, J. M., Houde, R. A., & Gayvert, R. T. (2006). Speech perception based on spectral peaks versus spectral shape. *The Journal of the Acoustical Society of America*, *119*(6), 4041-4054. https://doi.org/10.1121/1.2188369

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *20*(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597

Holliday, N. R., & Squires, L. (2020). Sociolinguistic labor, linguistic climate, and race(ism) on campus: Black college students' experiences with language at predominantly white institutions. *Journal of Sociolinguistics*, *25*(3), 418-437. https://doi.org/10.1111/josl.12438

Hollien, H., & Shipp, T. (1972). Speaking fundamental frequency and chronologic age in males. *Journal of Speech and Hearing Research*, *15*(1), 155-159. https://doi.org/10.1044/jshr.1501.155

Home Office. (2017). *Police and Criminal Evidence Act 1984 (PACE): Code D revised - Code of Practice for the identification of persons by police officers*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903812/pace-code-d-2017.pdf

House, A. S. (1960). Formant band widths and vowel preference. *Journal of Speech and Hearing Research*, *3*(1), 3-8. https://doi.org/10.1044/jshr.0301.03

Hughes, V. (2013). Establishing typicality: A closer look at individual formants. *Proceedings of Meetings on Acoustics* (pp. 1-9), Montreal, Canada. https://doi.org/10.1121/1.4798775

Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Forensic voice comparison using long-term acoustic measures of laryngeal voice quality.

Proceedings of the International Congress of Phonetic Sciences (pp. 1455-1459), Melbourne, Australia. https://assta.org/proceedings/ICPhS2019/

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & Segundo, E. S. (2018). The individual and the system: Assessing the stability of the output of a semi-automatic forensic voice comparison system. *Proceedings of Interspeech 2018* (pp. 227-231), Hyderabad, India. https://doi.org/10.21437/Interspeech.2018-1649

Ingram, J. C. L., Prandolinp, R., & Ong, S. (1996). Formant trajectories as indices of phonetic variation for speaker identification. *International Journal of Speech, Language and the Law*, *3*(1), 129-145. https://doi.org/10.1558/ijsll.v3i1.129

Ishikawa, K., & Webster, J. M. (2020). The formant bandwidth as a measure of vowel intelligibility in Dysphonic speech. *Journal of Voice, 37*(2). https://doi.org/10.1016/j.jvoice.2020.10.012

Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America*, *110*(2), 1141-1149. https://doi.org/10.1121/1.1384908

Jacewicz, E. (2005). Listener sensitivity to variations in the relative amplitude of vowel formants. *Acoustic Research Letters Online*, *6*(3), 118-124. https://doi.org/10.1121/1.1905384

Jain, K. C. (1979). A study of castes in Rajasthan (from 700-1200 A.D.). *Proceedings of the Indian History Congress* (pp. 144-149). https://www.jstor.org/stable/44141953

Jain, P. (2010). Bishnoi: An Eco-Theological “new religious movement” in the Indian desert. *Journal of Vaishnava Studies*, *19*(2), 1-20. https://ivsjournal.com/index.php/jvs/article/view/202

Jain, P. (2016). *Dharma and ecology of Hindu communities: Sustenance and sustainability* (2nd ed.). Routledge. https://doi.org/10.4324/9781315576916

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to statistical learning with applications in R* (2nd ed.). Springer Publication. https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *International Journal of Speech, Language and the Law, 4*(1), 84-103. https://doi.org/10.1558/ijsll.v4i1.84

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass, 2,* 671-711. https://doi.org/10.1111/j.1749-818X.2008.00066.x

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America, 108*(3), 1252-1252. https://doi.org/10.1121/1.1288413

Jung, J. W., Heo, H. S., Yang, I. H., Shim, H. J., & Yu, H. J. (2018). A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result. *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5349-5353), Calgary, AB, Canada. https://doi.org/10.1109/ICASSP.2018.8462575

Kahane, J. (1981). Anatomic and physiologic changes in the aging peripheral speech mechanism. In D. Beasley & G. Davis (Eds.), *Aging communication processes and disorders* (pp. 21-45). Grune and Stratton.

Kajarekar, S. S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L., & Bocklet, T. (2009). THE SRI NIST 2008 speaker recognition evaluation system. *Proceedings of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4205-4208), Taipei, Taiwan. https://doi.org/10.1109/ICASSP.2009.4960556

Kalmanovitch, Y., Leemann, A., Kolly, M. J., Schmid, S., & Dellwo, V. (2015). The role of longterm acquaintances in speech accommodation. In A. Leemann & M. Kolly, (Eds),

Trends in phonetics and phonology: Studies from German speaking Europe (pp. 93-107), Peter Lang.

Kardach, J., Wincowski, R., Metz, D. E., Schiavetti, N., Whitehead, R. L., & Hillenbrand, J. (2002). Preservation of place and manner cues during simultaneous communication: A spectral moments perspective. *Journal of Communication Disorders*, *35*(6), 533-542. https://doi.org/10.1016/S0021-9924(02)00121-1

Kassambara, A., & Patil, I. (2023*). ggcorrplot: Visualization of a correlation matrix using 'ggplot2'* (0.1.4.1 0.1.4.1) [Computer Software]. CRAN. https://CRAN.R-project.org/package=ggcorrplot

Kavoori, P.S. (2002). The Varna trophic system: An ecological theory of caste formation. *Economic and Political Weekly*, 37 (12), 1156-1164. http://www.jstor.org/stable/4411903

Kent, R. D., & Read, C. (2002). *The acoustic analysis of speech* (2nd ed., Vol. 10). Singular Thomson Learning.

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, *74*(November 2017), 74-97. https://doi.org/10.1016/j.jcomdis.2018.05.004

Kerswill, P. (2003). Dialect levelling and geographical diffusion in British English. In D. Britain & J. Cheshire (Eds.), *Social dialectology: In honour of Peter Trudgill* (pp. 223-243). John Benjamins Publishing Company. https://doi.org/10.1075/impact.16.16ker

Khera, R. (2005). *Drought proofing in Rajasthan: Imperatives, experience and prospects* (Discussion Papers Series, UNDP India). https://www.undp.org/sites/g/files/zskgke326/files/migration/in/drought_proofing_rajasthan_imperatives_experience_prospects.pdf

Kiefte, M., Enright, T., & Marshall, L. (2010). The role of formant amplitude in the perception of /i/ and /u/. *The Journal of the Acoustical Society of America*, *127*(4), 2611-2621. https://doi.org/10.1121/1.3353124

Kiesling, S. F. (2002). Men's identities and sociolinguistic variation: The case of fraternity men. *Journal of Sociolinguistics, 2*(1), 69-99. https://doi.org/10.1111/1467-9481.00031

Kinoshita, Y. (2001). *Testing realistic forensic speaker identification in Japanese: A likelihood ratio based approach using formants* [PhD Thesis, Australian National University]. Open Access Theses. http://hdl.handle.net/1885/110339

Kinoshita, Y., Osanai, T., & Clermont, F. (2022). Sub-band cepstral distance as an alternative to formants: Quantitative evidence from a forensic comparison experiment. *Journal of Phonetics, 94*, 101177. https://doi.org/10.1016/j.wocn.2022.101177

Klatt, D. H. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. *Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1278-1281). https://doi.org/10.1109/ICASSP.1982.1171512

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820-857. https://doi.org/10.1121/1.398894

Körkkö, P. (2015). Spectral moments analysis of/s/coarticulation development in Finnish-speaking children. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow. http://www.icphs2015.info/pdfs/Papers/ICPHS0470.pdf

Kothiyal, T. (2016). *Nomadic narratives: A history of mobility and identity in the great Indian desert*. Cambridge University Press.

Kothiyal, T. (2021). Frontiers, state and banditry in the Thar desert in the nineteenth century. In F. Ibrahim & T. Kothiyal (Eds.), *South Asian borderlands: Mobility, history, affect* (pp. 196-213). Cambridge University Press. https://doi.org/10.1017/9781108951500.009

Künzel, H. (1997). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech, Language and the Law*, *4*(1), 48-83. https://doi.org/10.1558/ijsll.v4i1.48

Künzel, H. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *International Journal of Speech, Language and the Law*, *8*(1). https://doi.org/10.1558/ijsll.v8i1.80

Labov, W. (1969). *A study of non-standard English*. ERIC Clearinghouse Linguistics. https://eric.ed.gov/?id=ED024053

Labov, W. (1972). Some principles of linguistic methodology. *Language in Society*, *1*(1), 97-120. https://doi.org/10.1017/S0047404500006576

Labov, W. (1973). *Sociolinguistic patterns* (No. 4). University of Pennsylvania Press.

Labov, W. (2006). *Social stratification of English in New York city*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618208

Ladefoged, P. (1996). *Elements of acoustic phonetics* (2nd ed.). The University of Chicago Press.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Wiley-Blackwell.

Ladefoged, P. (2006). *A course in phonetics* (5th ed.). Thomson Wadsworth Corporation.

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139166621

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014, 4-9 May 2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP) (pp. 1695-1699), Florence, Italy. https://doi.org/10.1109/ICASSP.2014.6853887

Lewis, M. P., & Summer Institute of Linguistics. (2009). *Ethnologue: Languages of the World* (M. P. Lewis, 16th ed.). SIL International. http://www.ethnologue.com

Li, L., Cheng, X., & Zheng, T. F. (2022). An application-oriented taxonomy on spoofing, disguise and countermeasures in speaker recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(2), e39. https://doi.org/10.1561/116.00000017

Li, X., Zhong, J., Wu, X., Yu, J., Liu, X., & Meng, H. (2020). Adversarial attacks on GMM i-vector based speaker verification systems. *Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6579-6583), Barcelona, Spain. https://doi.org/10.1109/ICASSP40776.2020.9053076

Lieberman, P., & Blumstein, S. E. (1988*). Speech physiology, speech perception, and acoustic phonetics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165952

Liepins, R., Kaider, A., Honeder, C., Auinger, A. B., Dahm, V., Riss, D., & Arnoldner, C. (2020). Formant frequency discrimination with a fine structure sound coding strategy for cochlear implants. *Hearing Research*, *392*, 1-9. https://doi.org/10.1016/j.heares.2020.107970

Lindau, M. (1978). Vowel features. *Language*, *54*(3), 541–563. https://doi.org/10.1353/lan.1978.0066

Lindblom, B., Diehl, R., & Creeger, C. (2009). Do ‘Dominant Frequencies’ explain the listener’s response to formant and spectrum shape variations? *Speech Communication*, *51*(7), 622-629. https://doi.org/10.1016/j.specom.2008.12.003

Lindblom, B., & Sundberg, J. (2014). The human voice in speech and singing. In: Rossing, T. (eds) *Springer handbook of acoustics*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-30425-0_16

Lindqvist-Gauffin, J., & Pauli. (1968). The role of relative spectrum levels in vowel perception. *Speech Transmission Laboratories (Stockholm) Quarterly Progress Report*, *9*(4), 012-015. http://www.speech.kth.se/qpsr

Liu, J. (2023). The development of Indo-European languages in cultural and historical processes: A case study of Indo-Iranian languages. Proceedings of *Advances in Social Science, Education and Humanities Research Proceedings of the 2nd International Conference on Humanities, Wisdom Education and Service Management (HWESM 2023)* (pp.16-21). https://doi.org/10.2991/978-2-38476-068-8_4

Lo, J. J. H. (2022). *fvclrr package: Likelihood ratio calculation and testing in forensic voice comparison* (1.1.4) [Package]. GitHub. https://github.com/justinjhlo/fvclrr/blob/master/README.md

Magier, D. S. (1983). *Topics in the grammar of Marwari* [PhD Thesis, University of California]. Berkeley ProQuest Dissertations Publishing. https://escholarship.org/uc/item/91q8z6pt

Maher, R. C. (2018). *Principles of forensic audio analysis* (Vol. 34). Springer International Publishing.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE* (pp. 561-580). https://doi.org/10.1109/PROC.1975.9792

Mandasari, M. I., McLaren, M., & Leeuwen, D. A. v. (2012). The effect of noise on modern automatic speaker recognition systems. *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP,)* (pp. 4249-4252), Kyoto, Japan. https://doi.org/10.1109/ICASSP.2012.6288857

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*(6), 3962-3973. https://doi.org/10.1121/1.2990715

Martin, P. (2021). *Speech acoustic analysis* (Vol. 1). Wiley Online Library. https://doi.org/10.1002/9781119808411

Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(2), 228-233. https://doi.org/10.1109/34.908974

Masica, C. P. (1991). *The Indo-Aryan languages.* Cambridge University Press.

McCandless, S. S. (1974). An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 22*(2), 135-141. https://doi.org/10.1109/TASSP.1974.1162559

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law, 11*(1), 103-130. https://doi.org/10.1558/sll.2004.11.1.103

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law, 13*(1), 89-126. https://doi.org/10.1558/ijsll.v13i1.89

McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. *Proceedings of the International Congress of Phonetic Sciences (ICPhS XVI)* (pp. 1825-1828), Saarbrücken, Germany. http://icphs2007.de/conference/Papers/1567/1567.pdf

Medabalimi, A. J. X., Seshadri, G., & Bayya, Y. (2014). Extraction of formant bandwidths using properties of group delay functions. *Speech Communication, 63-64*, 70-83. https://doi.org/10.1016/j.specom.2014.04.006

Meena, K. (2015). Diversity dimensions of India and their organization implications: An analysis. *International Journal of Economics & Management Sciences, 4*(6), 1-11. https://doi.org/10.4172/2162-6359.1000261

Meganathan, R. (2011). Language policy in education and the role of English in India: From library language to language of empowerment. In H. Coleman (Ed.), *Dreams and realities: Developing countries and the English language* (pp. 1-30). British Council UK. https://eric.ed.gov/?id=ED530679

Milenkovic, P., & Forrest, K. (1988). Classification of vowels using spectrum moments. *The Journal of the Acoustical Society of America*, *83*(S1), S67-S67. https://doi.org/10.1121/1.2025467

Miller, J. D. (1984). Auditory-perceptual correlates of the vowel. *The Journal of the Acoustical Society of America*, *76*(S1), S79-S80. https://doi.org/10.1121/1.2022028

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America*, *85*, 2114-2134. https://doi.org/ 10.1121/1.397862

Miller, R. L. (1953). Auditory tests with synthetic vowels. *The Journal of the Acoustical Society of America*, *25*(1), 114-121. https://doi.org/10.1121/1.1906983

Millhouse, T., Clermont, F., & Davis, P. (2002). Exploring the importance of formant bandwidths in the production of the singer's formant. *Proceedings of the 9th Australian International Conference of Speech Science and Technology* (pp. 1825-1828), Melbourne. http://icphs2007.de/conference/Papers/1567/1567.pdf

Mitra, V., Franco, H., Graciarena, M., & Mandal, A. (2012). Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4117-4120), Kyoto, Japan. https://doi.org/10.1109/ICASSP.2012.6288824

Mitsuya, T., MacDonald, E. N., Munhall, K. G., & Purcell, D. W. (2015). Formant compensation for auditory feedback with English vowels. *The Journal of the Acoustical Society of America*, *138*, 413–424. https://doi.org/10.1121/1.4923154

Moosmiiller, S. (1997). Phonological variation in speaker identification. *The International Journal of Speech, Language and the Law*, *4*(1), 29-47. https://doi.org/10.1558/ijsll.v4i1.29

Morrison, G. S. (2011a). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication, 53*(2), 242-256. https://doi.org/10.1016/j.specom.2010.09.005

Morrison, G. S. (2011b). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice, 51*(3), 91-98. https://doi.org/10.1016/j.scijus.2011.03.002

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., Ypma, R. J. F., Zhang, C., Anonymous, A., & Anonymous, B. (2021). Consensus on validation of forensic voice comparison. *Science and Justice*, *61*(3), 299-309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison, G. S., Sahito, F. H., Jardine, G. l., Djokic, D., Clavet, S., Berghs, S., & Dorny, C. G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, *263*, 92-100. https://doi.org/10.1016/j.forsciint.2016.03.044

Morrison, G. S., & Zhang, C. (2023). Forensic voice comparison – Overview. In Houck M., Wilson L., Eldridge H., Lewis S., Lothridge K., & Paul R. (Eds.), *Encyclopaedia of forensic sciences* (3rd ed., Vol. 2, pp. 737–750). Elsevier. https://doi.org/10.1016/B978-0-12-823677-2.00130-6

Mukherjee, K. (2011). Marwari. In Rajasthan: Part-I (pp. 29-142). Language Division Office of the Registrar General & Census Commissioner. https://www.academia.edu/44366931/MARWARI_KAKALI_MUKHERJEE

Nakatani, S. (2017). Hometowns of the Marwaris, diasporic traders in India. In S. Yamane & N. Naganawa (Eds.), *Regional routes, regional roots? Cross-border. patterns of human*

mobility in Eurasia (pp. 63-76). Hokkaido Slavic-Eurasian Research Center. https://www.ceeol.com/search/chapter-detail?id=561707

Nguyen, H. V., & Bai, L. (2011). Cosine similarity metric learning for face verification. In R. Kimmel, R. Klette, & A. Sugimoto (Eds.), *Computer Vision – ACCV 2010* (Vol. 6493, pp. 709-720). Springer. https://doi.org/10.1007/978-3-642-19309-5_55

Nigam, R. C. (1972). *Language handbook on mother tongues in Census*. Office of the Registrar General, Language Division, New Delhi.

Nijjar, B. S. (2008). The Jats (descendants of Scynthains). In *Origins and history of Jats and other allied nomadic tribes of India: 900 BC-1947 AD*. Atlantic Publishers & Dist.

Nissen, S. L., & Fox, R. A. (2005). Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective. *The Journal of the Acoustical Society of America*, *118*(4), 2570-2578. https://doi.org/10.1121/1.2010407

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *The Journal of the Acoustical Society of America*, *97*(1), 520-530. https://doi.org/10.1121/1.412278

Nolan, F. (1983). *The phonetic bases of speaker recognition* (1st ed.). Cambridge University Press.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *The International Journal of Speech, Language and the Law*, *12*(2), 143-173. https://doi.org/10.1558/sll.2005.12.2.143

Nolan, F., McDougall, K., & Hudson, T. (2011). Some acoustic correlates of perceived (dis)similarity between same-accent voices. *Proceedings of the International Congress of Phonetic Sciences (ICPhS XVII)* (pp. 1506-1509). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Nolan/Nolan.pdf

Office of the Registrar General & Census Commissioner. (2001). *Census of India, 2001: Language. India, states and union territories, Table C-16 (Vol. 16)*.

Office of the Registrar General & Census Commissioner. (2011a). *Census of India 2011 (15th Census)*. Retrieved from https://censusindia.gov.in/

Office of the Registrar General & Census Commissioner. (2011b). *Language : India, states and union territories (Table C-16)*. Retrieved from https://censusindia.gov.in/

Office of the Registrar General & Census Commissioner. (2011c). *Number of villages, towns, households, population and area (India, states/UTs, districts and Sub-districts)* (A-01). Retrieved from https://censusindia.gov.in/nada/index.php/catalog/42526

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E., (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures, *Journal of Memory and Language,69* (3), 183-195. https://doi.org/10.1016/j.jml.2013.06.002

Pardo, J. S., Pellegrino, E., Dellwo, V., & Möbius, B. (2022). Special issue: Vocal accommodation in speech communication. *Journal of Phonetics*, *95*, 101196. https://doi.org/10.1016/j.wocn.2022.101196

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence during conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1-11. https://doi.org/10.1016/j.wocn.2018.04.001

Patil, A., Gupta, C., & Rao, P. (2010). Evaluating vowel pronunciation quality: Formant space matching versus ASR confidence scoring. *Proceedings of the 2010 National Conference On Communications (NCC)* (pp. 1-5). https://doi.org/10.1109/NCC.2010.5430187

Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, *6*(61), 3167. https://doi.org/10.21105/joss.03167

Pellegrino, E., & Dellwo, V. (2023). Speakers are more cooperative and less individual when interacting in larger group sizes. *Frontiers in Psychology*, *14*, 1145572. https://doi.org/10.3389/fpsyg.2023.1145572

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175-184. https://doi.org/10.1121/1.1906875

Phillips, M. (2012). *Dialect continuum in the Bhil tribal belt: Grammatical aspects* [PhD Thesis, University of London]. http://eprints.soas.ac.uk/14048

Picheny, M., Durlach, N., & Braida, L. (1986). Speaking clearly for the hard of hearing. II Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, *29*(4), 434-445. https://doi.org/10.1044/jshr.2904.434

Potamianos, A., & Maragos, P. (1995). Speech formant frequency and bandwidth tracking using multiband energy demodulation. *The Journal of the Acoustical Society of America 99*(6), 3795-3806. https://doi.org/10.1121/1.414997

R Core Team. (2023). *R: A language and environment for statistical computing* (2023.06.1) [Computer Software]. R Foundation for Statistical Computing. https://www.R-project.org/

R v Slade & Ors [2015] EWCA Crim 71, paras 43-45 (Court of Appeal ,Criminal Division). http://www.bailii.org/ew/cases/EWCA/Crim/2015/71.html

Ram, D. (2014). Sex and residence-wise analysis of literacy in Rajasthan. *Journal of Research in Humanities and Social Science*, *2*(3), 23-28. https://www.questjournals.org/jrhss/papers/vol2-issue3/C232328.pdf

Ramig, L. A., & Ringel, R. L. (1983). Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech and Hearing Research*, *26*(1), 22-30. https://doi.org/10.1044/jshr.2601.22

Rathore, L. S., & Saxena, K. S. (1987). Politics and caste in Rajasthan. *The Indian Journal of Political Science*, *48*(4), 449-457. https://www.jstor.org/stable/41855330

Reddy, N. S., & Swamy, M. N. S. (1984). High-resolution formant extraction from linear-prediction phase spectra. *Proceedings of IEEE Transactions on Acoustics, Speech, and Signal Processing* (pp. 1136-1144). https://doi.org/10.1109/TASSP.1984.1164456

Reubold, U., Harrington, J., & Kleber, F. (2010). Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication*, *52*(7-8), 638-651. https://doi.org/10.1016/j.specom.2010.02.012

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. *Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. IV-4072-IV-4075), Orlando, FL, USA. https://doi.org/10.1109/ICASSP.2002.5745552

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1-3), 19-41. https://doi.org/10.1006/dspr.1999.0361

Rhodes, R. (2017). Aging effects on voice features used in forensic speaker comparison. *The International Journal of Speech, Language and the Law*, *24*(2), 177-199. https://doi.org/10.1558/ijsll.34096

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2023). *Support functions and datasets for Venables and Ripley's MASS* (7.3-60) [Computer Software]. CRAN R. http://www.stats.ox.ac.uk/pub/MASS4/

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2016). *Interpreting evidence: Evaluating forensic science in the courtroom* (2nd ed.). John Wiley & Sons.

Rodman, R., McAllister, D., Bitzer, D., Cepeda, L., & Abbitt, P. (2002). Forensic speaker identification based on spectral moments. *The International Journal of Speech, Language and Law*, *9*(1), 22–43. https://doi.org/10.1558/sll.2002.9.1.22

Rose, P. (2002). *Forensic speaker identification* (1st ed.). CRC Press. https://doi.org/10.1201/9780203166369

Rossing, T. D. (2014). *Springer handbook of acoustics*. Springer.

Roy, T. (2015). Diaspora: Marwari. In *Oxford Handbook Topics in History*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199935369.013.22

Sadiq, S., & Harwardt, C. (2011). Investigations on the speech spectrum of normal and high-effort speech. *Proceedings of Interspecch 2011* (pp. 2937-2940), Florence, Italy. https://doi.org/10.21437/Interspeech.2011-735

Saha, A. (1993). The caste system in India and its consequences. *International Journal of Sociology and Social Policy, 13*(3/4), 1-76. https://doi.org/10.1108/eb013170

Sakayori, S., Kitama, T., Chimoto, S., Qin, L., & Sato, Y. (2002). Critical spectral regions for vowel identification. *Neuroscience Research*, *43*(2), 155-162. https://doi.org/10.1016/S0168-0102(02)00026-3

Samuvel, N., Joshua, M., Koshy, B., & Abraham, B. (2012). *Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India: Preliminary overview* (SIL Electronic Survey Reports 2012-029). https://www.sil.org/system/files/reapdata/16/69/61/166961978548220170848633571549700153259/silesr2012_029.pdf

Savela, J., Ojala, S., Aaltonen, O., & Salakoski, T. (2007). Role of different spectral attributes in vowel categorization: The case of Udmurt. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)* (pp. 384-388), Tartu, Estonia. https://aclanthology.org/W07-2462

Schindler, C., & Draxler, C. (2013). Using spectral moments as a speaker specific feature in nasals and fricatives. *Proceedings of Interspeech 2013* (pp. 2793-2796), Lyon, France. https://doi.org/10.21437/Interspeech.2013-639

Schwartz, J. C., Whyte, A. T., Al-Nuaimi, M., & Donai, J. J. (2018). Effects of signal bandwidth and noise on individual speaker identification. *The Journal of the Acoustical Society of America*, *144*(5), EL447-EL452. https://doi.org/10.1121/1.5078770

Senoussaoui, M., Kenny, P., Dumouchel, P., & Dehak, N. (2013). New cosine similarity scorings to implement gender-independent speaker verification. *Proceedings of Interspeech 2013* (pp. 2773-2777), Lyon, France. https://doi.org/10.21437/Interspeech.2013-635

Shougrakpam, D. (2022). Language, A representation of cultural heritage and identity-Marwari, Rajasthan, India. *Journal of Asian Arts, Culture and Literature (JAACL)*, *3*(2), 1-8. https://alsphere.org/wp-content/uploads/2022/06/Language-A-Representation-of-Cultural-Heritage-and-Identity-Marwari-Rajasthan-India-by-Dr.-Dhanapati-Shougrakpam-1.pdf

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and linguistics compass*, *3*(2), 621-640. https://doi.org/10.1111/j.1749-818X.2009.00125.x

Singh, S. K., & Pandey, P. C. (2003). *Features and techniques for speaker recognition* (M. Tech. Credit Seminar Report). https://www.ee.iitb.ac.in/~esgroup/es_mtech03_sem/sem03_paper_03307409.pdf