university of groningen

campus fryslân

Speaking Volumes: How Acoustic Features Reveal Speaker Height

Stella Siu

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Speaking Volumes: How Acoustic Features Reveal Speaker Height

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Matt Coler (Voice Technology, University of Groningen)

Stella Siu (4052455)

July 9, 2025

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Matt Coler, for his unwavering support, insightful guidance, and generous time. His care and encouragement throughout this journey have been invaluable, and I am truly grateful for the opportunity to learn under his mentorship.

I would also like to sincerely thank Vass and Hieke for their kind and supportive help, especially during the difficult period. Their understanding and compassion meant a great deal to me and helped me stay strong and continue my thesis writing.

My heartfelt thanks go to the VT team—Phat, Joshua, and Shekhar—for their exceptional dedication and passion for teaching. When I was unable to attend some classes due to health issues, their willingness to offer immediate extra support and assistance allowed me to catch up and stay on track. Their generosity and commitment left a lasting impression on me.

To my wonderful classmates—Jan, Tiantian, Max, Jiashu, Monica, and Millie—thank you for your friendship, encouragement, and the inspiring moments we shared. Your presence made this year not only academically fulfilling but also personally unforgettable. I will always cherish how we helped one another and grew together.

I am profoundly thankful to my family for their boundless love and support. I especially want to acknowledge my 101-year-old grandmother, who raised me and always loves me, and my younger brother, whose guidance in programming was an incredible help throughout my studies. I also extend my thanks to my best friends in Hong Kong and Japan—you have never been far from my thoughts, and your support has been felt across the distance.

Above all, I want to thank Franz for being my rock and standing by me through all the ups and downs. Your love, patience, and encouragement have meant everything to me. I also thank my beloved cats—Coffee, Cola, and Nimbus—for their comforting presence and companionship that brought light to many challenging days.

To those who never truly existed, yet never truly left—your strength became mine.

Abstract

With growing interest in biometric technologies, speaker height estimation directly from acoustic signals has emerged as a valuable capability for applications in forensics, authentication, and speech profiling. However, most state-of-the-art systems rely on full speech input, which poses challenges for conversational privacy. This study investigates the feasibility of predicting speaker height from sub-lexical acoustic features using lightweight models. Basic feature (F0), intermediate features (formants), and high-dimensional features (MFCCs) were utilized as input across three regression models: simple linear, multiple linear, and random forest regression.

Results show that MFCCs combined with multiple linear regression yield a statistically significant performance using only isolated diphthong /aw/, achieving a minimum root-mean-square error (RMSE) below 7 cm on the TIMIT dataset. This performance is on par with state-of-the-art full speech input and deep neural network models. MFCCs also showed greater gains when used with multivariate models, suggesting that feature complexity and model structure interact to influence prediction outcomes. Additionally, diphthong /aw/ emerged as the most reliable input unit, consistently yielding low prediction errors in both multiple linear and random forest regressions, whereas reduced vowel /ax-h/ consistently underperformed across all feature sets and regression models. Furthermore, an inverse relationship between F1 and F4 was observed in both simple linear regression and random forest feature importance analysis, indicating that as one becomes more predictive, the other contributes less—suggesting a complementary dynamic in height estimation.

These findings demonstrate that phone based input, which is linguistically impoverished, can reduce conversational privacy risks and offer a viable alternative to models based on full speech. They suggest a promising direction for developing interpretable and conversational privacy conscious speaker profiling systems using minimal speech input.

Contents

1 Introduction

Speech is central to human communication, enabling the exchange of ideas, emotions, and intentions. Yet beyond conveying linguistic and paralinguistic content, speech also encodes information about the speaker. For example, it reveals demographic attributes such as age and gender (Schilling & Marsters, 2015), as well as physical characteristics like vocal tract length (Lammert & Narayanan, 2015). Together, these cues contribute to a distinctive vocal signature, making speech both a communicative tool and a biometric marker. This means that speech is not only a tool for linguistic communication but also a medium revealing speaker-specific traits. Because of this dual function, speech can serve as a powerful input for various speech technologies, particularly in the field of Automatic Speech Recognition (ASR), which seeks to convert spoken language into written text. While the primary goal of ASR is accurate word recognition, this field has also been extending the applications by increasingly integrating speaker-related information to adapt to diverse speaking populations, for example, the domains of speaker verification using speaker recognition and speaker profiling for communities with different background. Beyond speaker identification and verification, ASR systems can extract demographic and physiological traits through speaker profiling. Speaker profiling involves extracting speech and inferring stable physiological and demographic traits, such as age, gender, and height, using acoustic features. To improve accuracy in estimating height, state-of-the-art research has increasingly looked beyond traditional features, and rather converted full speech into vectors as input of height estimation tasks through deep neural network models (Poorjam, Bahari, Vasilakakis, & Hamme, 2015; Rajaa, Van Tung, & Siong, 2021).

On one hand, the use of full speech signal making height prediction feasible objectively, for example Rajaa et al. (2021)'s single-task setting for height prediction have achieved a root mean square error (RMSE) of 6.0 for female speakers. This encourages the technological trend of widely using biometric voice recognition technologies for identity management (Mohammed & Ali, 2024). For instance, financial institutions such as HSBC employ VoiceID as a unique vocal identifier for secure telephone banking access. In the educational context, voice recognition systems are increasingly used in online assessments to authenticate users, particularly individuals with disabilities (Rudrapal, Das, Debbarma, Kar, & Debbarma, 2012), help prevent identity fraud, such as impersonation by a "ringer" (Yee & MacKown, 2009), and uphold academic integrity (Hernandez-de Menendez, Morales-Menendez, Escobar, & Arinez, 2021).

On the other hand, embeddings derived from the entire speech signal raise both biometric privacy and conversational privacy concerns. First, although using the full speech signal for prediction and estimation is highly effective, it necessarily involves collecting sensitive biometric data. If this information is misused or inadequately stored, it can lead to serious security risks. Second, full speech recordings inherently capture semantic content that may include sensitive personal information, making people wary of systems that store or analyze complete utterances. This apprehension is reflected in growing public awareness and mistrust toward home devices perceived to "listen in" on conversations. Interestingly, many users are less concerned about their use of biometric data than about the risk of conversational content being monitored or repurposed Despres et al. (2024). This raises a need to research methods that use linguistically impoverished input, such as sub-lexical segments, to achieve predictive performance comparable to models trained on full speech recordings. Therefore, developing high-performing models that can achieve comparable results using minimal

input data, for example, phone-based speech utterances, could strengthen public confidence in biometric voice recognition technologies and encourage broader acceptance by alleviating concerns about conversational privacy.

Research on speaker profiling, especially height estimation, has a long history that predates the use of deep neural networks and the use of full speech signals. These earlier studies based on acoustic-physiological mechanism may offer valuable insights and directions for contemporary research on height prediction from speech aimed at preserving speaker privacy.

1.1 Acoustic-physiological Mechanism

Early studies investigating the relationship between vocal tract length (VTL), formant frequencies, and speakers' height can be traced back to the twentieth century. Based on the source-filter theory of speech (Fant, 1960), the vocal tract acts as an acoustic filter, with its length determining the spacing between resonances (formant frequencies) of supralaryngeal vocal-tract. A longer vocal tract results in lower and more closely spaced formant frequencies. Building on this theoretical framework, Fitch and Giedd (1999) demonstrated a strong correlation between speaker height and vocal tract length (VTL) across individuals of varying ages and statures ($r = 0.93$), with taller speakers typically exhibiting longer vocal tracts. Along with (Fant, 1960), Fitch and Giedd (1999) established a link between VTL, formant frequencies, and speaker height, suggesting that taller individuals tend to have longer VTLs and consequently produce speech with lower formant frequencies, and vice versa. These observations imply that the relationship between speaker height and acoustic features is mediated by systematic anatomical variations (González, 2004; Lass & Davis, 1976). Given this relationship, acoustic features may serve as a viable basis for predicting speaker height. Nonetheless, empirical results have been inconsistent. Several studies have reported limited or unreliable correlations, and most attempts to accurately estimate height from formants have not been successful (González, 2004; Hatano et al., 2012; Lammert & Narayanan, 2015).

Given the inconsistent correlations between formants and height, researchers have explored alternative acoustic features for height estimation. Features such as fundamental frequency (F0) and Mel-frequency cepstral coefficients (MFCCs) have gained attention for their potential to capture subtle physiological differences related to height (Dusan, 2005; Ganchev, Mporas, & Fakotakis, 2010). Although F0 is more directly influenced by vocal fold properties, it may reflect trends associated with speaker size and may correlate with physical dimensions. MFCCs, as compact representations of the speech spectrum, are widely used in speech processing. MFCCs have been assessed the correlation with height estimation (Dusan, 2005) and have also been used as input of deep neural networks for height prediction (Rajaa et al., 2021).

However, different acoustic features capture these height-related variations with varying effectiveness, as evidenced by the contradictory findings reported by Dusan (2005) and Ganchev et al. (2010).

1.2 Research Gap

According to Dusan (2005)'s findings, using multi-linear regression, higher-dimensional acoustic features such as MFCCs, were shown to exhibit stronger correlations with speaker height (r = 0.74), whereas lower-dimensional features, such as fundamental frequency (F0), showed weaker correlations (r = 0.59). This suggests that higher-dimensional features may be more informative and accurate for predicting height. This view is further supported by previous studies that relied on lower-level features such as formants, which often reported low correlations with speaker height (González, 2004; Hatano et al., 2012). When Ganchev et al. (2010) employed openSMILE to extract and rank acoustic features for height prediction based on prediction error, although MFCCs remained the top-performing features, F0 also appeared multiple times among the top 50 features, which contrasted with Dusan (2005)'s findings and proved that lower-dimensional features may still carry numeral relevant information for height estimation.

Therefore, the first research gap concerns whether correlation translates to prediction accuracy in height prediction. Dusan (2005)'s study assumed higher dimension features are more useful to predict height without demonstrating that increasing the dimensionality of acoustic features leads to actually improved performance as measured by lower prediction error. This assumption was challenged by the research of Ganchev et al. (2010), which assessed the effectiveness of acoustic feature subsets in height prediction using prediction error metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Ganchev et al. (2010) particularly found that three subsets of F0 features ranked in the top 10, despite F0 being the acoustic feature with lower correlation in Dusan (2005)'s study. This rationale underpins the choice of RMSE as an appropriate evaluation metric for this study and motivates the development of Sub-Research Question 1 (Sub-RQ1).

Furthermore, the second research gap concerns phone-specific patterns that have been largely overlooked in prior work. Dusan (2005)'s correlation results differed significantly between vowels, the Pearson correlation coefficient for /iy/ (r = 0.73) was approximately double that of the reduced vowel /ax-h/ (r = 0.36), highlighting significant vowel-dependent differences in height predictability. However, it remains unclear whether these vowels are also associated with higher prediction accuracy. This observation highlights the need for further analysis of vowel-specific performance measured by prediction error and points to the potential utility of phone-level modelling for height prediction, leading to the formulation of Sub-Research Question 2 (Sub-RQ2).

The third research gap concerns the comparison between linear and non-linear regression models for height prediction. Dusan (2005) used a multiple linear regression model, while Ganchev et al. (2010) applied a Support Vector Machine (SVM) with a kernel function, representing a non-linear approach. This methodological difference may partly account for the discrepancies in their findings and underscores the need to examine the performance of both linear and non-linear regression models using the same set of acoustic features. The goal is to determine which approach more effectively captures the relationship between acoustic features and speaker height. In addition, this study uses feature importance scores from the random forest model to investigate specific relationships between acoustic features and height, contributing to the formulation of Sub-Research Questions 3 and 4 (Sub-RQ3 and Sub-RQ4).

1.3 Research Questions and Hypotheses

In light of the preceding discussion on the identified research gaps, and based on the phone-level approach used by Dusan (2005), which suggests a valuable direction for height prediction using minimal speech input and aligns with privacy considerations, I propose a tiered analysis that investigates how acoustic feature complexity and model choice affect the accuracy of speaker height estimation at the phone level, with a focus on vowels. In this study, the terms vowel and phone will used interchangeably. The rationale for this will be further explained in 3.2. This study addresses the following research question:

> **How does acoustic feature complexity affect height prediction accuracy when comparing basic features (F0), intermediate features (formants), and high-dimensional features (MFCCs) across different regression models at phone-based level?**

This main question can be broken down into the following sub-questions:

1. How does the use of different feature sets (basic, intermediate, high-dimensional) impact RMSE across phones and regression models? In particular, do high-dimensional features (MFCCs) consistently outperform simpler features?

2. Are there specific phones for which height can be predicted most or least accurately, and do these patterns align with articulatory openness (open vs. closed vowels) and phonetic reduction?

3. How does the performance (RMSE) of linear regression models (simple and multiple) compare to that of a non-linear model (random forest regression) in predicting speaker height from acoustic features across phones?

4. Which acoustic features contribute most significantly to height prediction in the random forest regression model, and how do feature importance patterns relate to the acoustic-physiological mechanisms?

- H1: High-dimensional acoustic features (MFCCs) will produce lower RMSE values than basic (F0) and intermediate features (formants) across most phones when using linear regression models. This hypothesis is inspired by Dusan (2005), who reported that the correlation between speaker height and acoustic features increases with the dimensionality of those features using multiple linear regression model. However, this hypothesis aims to explore under what conditions different feature sets most effectively predict height, rather than to assert a uniform superiority of MFCCs across all linear modelling scenarios.

- H2: The phone /ax-h/ will consistently exhibit the highest RMSE across all feature sets (Dusan, 2005), due to its status as a reduced vowel characterized by high articulatory and acoustic variability. Open vowels (/aa/, /ae/, /aw/) are expected to yield lower RMSE values, as they involve greater vocal tract expansion compared to close vowels. This stretching enhances the acoustic distinction between speakers of different heights, leading to more consistent cues for height prediction.

- H3: Multiple linear regression will outperform simple linear regression, but the improvement will be more pronounced for high-dimensional features (MFCCs with 13 dimensions) than for intermediate features (formants with four dimensions) (Dusan, 2005).

To validate the proposed hypotheses, RMSE values obtained from the simple linear regression, and random forest regression using formant and MFCC features are first aggregated by computing the mean RMSE across all features for each phone. This results in one representative RMSE value per phone, allowing for appropriate statistical comparison. Table 1 below presents the hypotheses falsifiability criteria:

Hypothesis	**Test**	**IV**	**DV**	**Accepted if:**
H1	Friedman	F0, Formants, MFCCs (from SR)	(Mean) RMSEs of phones	It is considered validated if statistically significant results ($p < 0.05$) allow analysis of the conditions under which each feature set predicts height most effectively.
	Wilcoxon	Formants, MFCCs (from MR)	RMSEs of phones	
H2	Friedman	20 Phones	RMSE of SR	/ax-h/ consistently yields higher RMSEs AND one of the open vowels (/aa/, /ae/, /aw/) consistently yields lower RMSEs across all 3 tests
	Wilcoxon	20 Phones	RMSEs of MR	
	Wilcoxon	20 Phones	RMSEs of RF	
H3	Wilcoxon	SR, MR	Mean RMSEs of Formants	MR outperform SR in both tests
	Wilcoxon	SR, MR	Mean RMSEs of MFCCs	
	Wilcoxon	Formants, MFCCs	ΔRMSE	Improvement of MFCCs is more pronounced than of formants

Table 1: Overview of Hypotheses Validation

1.4 Thesis Outline

Now that the motivation for this research has been presented, the structure of this thesis is as follows:

- Section 2 reviews relevant literature categorized into three main themes, and then situates the present study within a research framework developed around the gap highlighted in the works of Dusan (2005) and Ganchev et al. (2010).
- Section 3 describes the dataset, acoustic features selected, regression models, evaluation methodology and ethical considerations of this work.

- Section 4 details the data preparation steps and experiment setups.
- Section 5 presents and analyzes the regression results and validates the hypotheses using statistical tests.
- Section 6 discusses key insights, addresses limitations, and provides answers to the research questions.
- Section 7 concludes with key findings and future directions.

2 Literature Review

This chapter begins by outlining the search strategy and selection criteria used in the literature review process. Following this, a comprehensive review of prior research on height prediction using speech signals is presented. The review is organized into three thematic categories, followed by a systematic comparative analysis of research gaps identified in the studies by Dusan (2005) and Ganchev et al. (2010):

- **Perception Studies:** Research focused on human listeners' perception of apparent height based on speech stimuli.
- **Correlation Studies:** Studies investigating the relationship between height or vocal tract length (VTL) and one or more sets of acoustic features.
- **State-of-the-Art (SOTA) Performance:** Research employing advanced machine learning techniques, including deep neural networks, for height prediction from speech.
- **Research Gap Analysis:** A detailed examination of the limitations and gaps in the works of Dusan (2005) and Ganchev et al. (2010).

2.1 Search Strategy and Selection Criteria

The literature search strategy employed a combination of Boolean operators and quoted search terms on Google Scholar, arXiv, and IEEE, using targeted keywords to gather a diverse and relevant set of sources. Keywords from each thematic category were incorporated in every search to ensure comprehensive coverage.

Primary search:

- **Main Theme:** "height estimation" OR "speaker height prediction" OR "physical traits from speech"
- **Acoustic Features:** "formant frequencies" OR "F0" OR "F1" OR "F2" OR "F3" OR "F4" OR "MFCC" OR "acoustic features" OR "speech signal"
- **Machine Learning Method:** "machine learning" OR "deep learning" OR "neural network"

Secondary search:

- **Main Theme:** "biometric voice recognition" OR "speaker profiling" OR "voice biometrics"
- **Acoustic Features:** "formant frequencies" OR "F0" OR "F1" OR "F2" OR "F3" OR "F4" OR "MFCC" OR "acoustic features" OR "speech signal"

Explanation of selection criteria:

1. Inclusion criteria: Peer-reviewed journal articles and conference papers reporting empirical results, both subjective and objective evaluation, related to height prediction from speech using acoustic features or full speech signal as input.
2. Exclusion criteria: Studies focusing on VTL structure analysis and acoustic features only.

2.2 Perception Studies

To begin with, the literature on human perception of a speaker's height from speech includes two key perceptual studies conducted by the same group of researcher at different times. Although the present study does not aim to investigate human listeners' subjective estimations of VTL or speaker's height, previous work by Barreda (2016) and Barreda and Predeck (2024) still provides useful context for understanding the complexity of height prediction from speech based on vowel acoustics.

Formant values and vowel identity could mislead listeners' height judgments systematically. When vowels implied similar VTLs, the back vowel /ʊ/ was perceived to be "taller" than the front vowels /i/ and /ae/, likely due to its lower F1 and F2 values (Barreda, 2016). On the other hand, the inclusion of higher formants improved perceptual accuracy, suggesting that listeners relied not only on VTL-related cues but also on vowel-specific spectral characteristics (Barreda, 2016). The significant influence of vowel identity on perceived height is further supported by Dusan (2005), whose findings revealed varying correlation coefficients between vowels, acoustic features, and speaker height. This indicates that the effectiveness of vowels and acoustic features in height prediction models can differ substantially in terms of accuracy. These differences warrant closer analysis through phone-specific patterns, highlighting a research gap concerning phoneme-specific effects in height prediction. In particular, there is a need to examine feature importance scores derived from random forest regression across formant frequencies, especially F1, F2, and F4, to assess whether higher formants offer supplementary cues that align with human perceptual strategies.

Furthermore, Barreda and Predeck (2024) argued that human listeners had an "underlying systematic process" and used social knowledge to produce stable speaker height judgments without needing higher formant information, but this conversely suggests that regression models which lacking such cognitive mechanisms, may yield results that do not reflect a similarly systematic process. This also underscores the importance of examining feature importance scores from random forest regression to identify which acoustic cues drive model predictions.

2.3 Correlation Studies

Fant (1960)'s source-filter theory of speech provided the theoretical basis for understanding the correlation between vocal tract length (VTL) and formant frequencies. Building on this, Fitch and Giedd (1999) employed magnetic resonance imaging (MRI) to examine the relationship between VTL and body size, offering direct anatomical evidence. Fitch and Giedd (1999) reported a strong correlation between speaker height and VTL ($r = 0.93$), demonstrating the potential of VTL as a predictor of height. Together, these findings suggest that taller individuals tend to have longer VTLs, which in turn produce lower formant frequencies. Conversely, formant frequencies can serve as indirect indicators of VTL, making it possible to infer a speaker's height. These studies are critically important, as they establish a fundamental connection between VTL, formant structure, and speaker height—forming the basis for height estimation using acoustic cues.

Nonetheless, one critical limitation of Fitch and Giedd (1999)'s study is that subsequent attempts to replicate or extend its correlation findings have often failed to yield consistent or reliable results. For example, both González (2004) and Hatano et al. (2012) aimed to extend the earlier findings

from English to other languages—Spanish and Japanese respectively. However, none of the F0 coefficients reached statistical significance, and all correlation values were below 0.60 in absolute terms (González, 2004). Similarly, Hatano et al. (2012) reported that VTL did not reliably predict body height, F0, or formant frequencies in adult male speakers. Therefore, F0 and formants, as basic and intermediate-dimensional acoustic features, can be seen as unreliable predictors of speaker height according to the subsequent studies. However, it is important to note that Hatano et al. (2012) was significantly constrained by a small sample size—only five Japanese speakers—and a limited set of five vowels, underscoring the broader issue of insufficient data in this line of research.

On the other hand, Lammert and Narayanan (2015) also examined the relationship between VTL and formant frequencies using data from five speakers, but evaluated their results using RMSE rather than correlation coefficients. Their findings were insightful: the results generally supported Fitch and Giedd (1999), and found that the lowest RMSE values were associated with higher formants, which are less influenced by articulation (Lammert & Narayanan, 2015). Higher formant values may support more accurate VTL estimation and speaker height inference, aligning with patterns observed in human perceptual judgments reported by Barreda (2016). Along with the findings of González (2004), which showed that F2 of /e/ exhibited a relatively strong correlation with height for both male and female speakers, these results collectively underscore the need to investigate phone-specific patterns in height estimation, particularly in light of the varied and inconsistent conclusions across studies.

Nonetheless, the aforementioned studies are limited to using only formant frequencies or combinations of F0 and formants as acoustic features, without directly examining the relationship between speaker height, VTL, and a broader set of acoustic features. Addressing this gap, Dusan (2005) investigated various acoustic feature sets at the vowel level and their correlations with speaker height. The study yielded two key insights. First, higher correlation values between acoustic features and height were associated with increased dimensionality, as they accounted for a greater proportion of variability in speaker height. Reported correlation coefficients were MFCC (r = 0.74), LPC (r = 0.73), formants (r = 0.73), and F0 (r = 0.59) (Dusan, 2005), suggesting that higher-dimensional features may be more informative for height prediction when considered individually. The conclusion of the meta-analysis by Pisanski et al. (2014) further supported Dusan (2005) with the analysis that F0 accounted for a maximum of 2% of the variance in human height, while formants explained up to 10% of height variation within sexes. Second, a combined feature set comprising MFCCs, LPCs, and formants could account for 57.2% of the variability in speaker height. This finding positions MFCCs as a preferred input over other feature types in SOTA experiments, due to their strong correlation with speaker characteristics. It also motivated the use of full speech signals as input, given that full speech is even higher in dimensionality than MFCCs and may therefore capture more speaker height variability. However, this study has several limitations, which, along with the contradictory findings reported by Ganchev et al. (2010), will be discussed in Section 2.5 as the basis for identifying the main research gap.

In short, correlation studies have provided valuable insights into how different acoustic features vary in their association with speaker height, guiding the selection of input features for height prediction models, but they assess mainly the theoretical usefulness of features based on correlation, without evaluating actual model performance through prediction error metrics such as RMSE. While

perception studies have provided further insight into phone-specific feature relevance on top of correlation studies, they typically examine only five vowels, resulting in insufficient data to analyze why certain phones may be more informative. Additionally, both types of studies have primarily relied on limited regression approaches, without exploring how different regression models might influence results. These limitations collectively point to the three key research gaps addressed in this study.

2.4 State-of-the-Art (SOTA) Performance

State-of-the-art (SOTA) studies typically follow three key practices: (1) employing advanced machine learning techniques to extract features from full speech signals, (2) utilizing deep neural networks as the estimation models, and (3) training and evaluating their systems using existing datasets with annotated metadata (Poorjam et al., 2015; Rajaa et al., 2021). The primary goal of these SOTA approaches is to estimate or predict speaker height with improved accuracy, aiming to minimize prediction error.

As noted in Section 2.3, SOTA architectures commonly use MFCCs either as input features or as components of embedded representations. For instance, both Poorjam et al. (2015) and Kalluri, Vijayasenan, and Ganapathy (2019) extracted 20 MFCCs, with Poorjam et al. (2015) expanding them into a 60-dimensional vector by including first and second-order derivatives, while Kalluri et al. (2019) used MFCCs only in a baseline system. In contrast, Rajaa et al. (2021) did not use MFCCs directly, opting instead for an unsupervised encoder that processes full speech signals. Since these studies employed advanced embeddings and sophisticated models, they did not place emphasis on understanding or analyzing the input features but only evaluate the prediction performance. From a privacy-conscious perspective, using minimal yet informative input features combined with a complex model may represent a promising direction for future research. Building on these studies, the use of RMSE remains a suitable evaluation metric for assessing prediction performance. Across these approaches, SOTA systems typically report best-case RMSEs in the range of 6 to 7 cm for height estimation. Therefore, in this study, any RMSE result falling within the 6 to 7 cm range can be considered outstanding to be proposed as good minimal input, as it is achieved without the use of advanced machine learning architectures.

2.5 Research Gap Analysis

Section 2.3 discussed the findings of Dusan (2005), which suggest that higher-dimensional acoustic features may be more informative for height prediction. This implies an underlying assumption that higher-dimensional features are more effective, meaning more accurate, for predicting height. This notion is further supported by the findings of González (2004) and Hatano et al. (2012), which show that basic and intermediate features such as F0 and formants yield low correlation values in height prediction. While Ganchev et al. (2010) acknowledged the strong relevance of MFCCs, noting that half of the top 50 parameters were statistical functionals derived from them, F0 also appeared multiple times among the top-ranked features. This latter finding challenges the underlying assumption of a limited role for F0 in height prediction and suggests that correlation alone does not necessarily translate to predictive accuracy when using acoustic features. Therefore, the first question arises: Does the prediction performance align more closely with the correlation-based findings of Dusan (2005) or with the audio feature ranking results reported by Ganchev et al. (2010)? This question

motivates sub-RQ1, as outlined in 1.3.

Dusan (2005) reported that close and mid-close front vowels, showed stronger correlations between speaker height and MFCC features, for example, /iy/ (r = 0.73), /ih/ (r = 0.70), and /ix/ (r = 0.69). While this provides substantial evidence that correlation strength varies across vowels in multiple regression, it remains unclear whether these vowels actually yield higher prediction accuracy. The original study focused primarily on acoustic feature sets, without further analysis of vowel identity. This represents a significant research opportunity, especially given insights from Sections 2.2 and 2.3, which suggest that vowel identity plays an important role in height perception and estimation. From a phonetic perspective, vowel distinctions are shaped by modifications in vocal tract configuration that primarily affect the first three formant frequencies, but, higher formants such as F4 and F5 additionally contribute to voice projection and resonance (Story, 2004), which actually aligned with Barreda (2016)'s findings. Given this physiological basis and its potential implications for speaker profiling, further exploration of vowel-specific effects on height estimation is both justified and promising to formulate sub-RQ2 as outlined in 1.3.

One of the key differences between Dusan (2005) and Ganchev et al. (2010) lies in their choice of modeling approach: the former used multiple linear regression, while the latter applied a Support Vector Machine (SVM) with a kernel function, representing a non-linear method. This methodological contrast may explain the inconsistencies in their findings. Moreover, neither study thoroughly investigated the interaction between acoustic features, vowel identity, and speaker height. Instead, they focused separately on feature-height correlation or predictive performance, overlooking the critical role that vowel identity may play in shaping these relationships. Notably, this gap can be explored using feature importance scores from random forest models, which are well-suited for uncovering such patterns. Therefore, the present study addresses this limitation by evaluating prediction accuracy across different regression models, while also examining the influence of vowel identity and feature types through detailed analysis of random forest feature importance.

This section identifies three key research gaps: (1) the assumption that higher-dimensional features improve height prediction is untested, as correlation does not guarantee accuracy—highlighted by conflicting findings from Dusan (2005) and Ganchev et al. (2010); (2) vowel identity's role in height estimation remains unexplored despite evidence showing variation in feature correlations across vowels; and (3) prior studies used differing regression models without examining how model type affects performance. This study addresses these gaps by comparing feature sets, evaluating both linear and nonlinear models, and analyzing vowel-specific patterns using random forest feature importance to inform sub-research questions.

Table 2: Summary of Key Literature

Reference	Key Findings	Theme
Barreda (2016)	Investigated how formant frequencies influence listener judgments of speaker's height, suggesting human perceived different height from vowel quality with similar VTLs, and suggested higher formants improved perceptual accuracy.	Perception Studies
Barreda and Predeck (2024)	Suggested human listeners had an "underlying systematic process" and would use social knowledge to estimate speaker's height.	Perception Studies
Fant (1960)	Proposed the source-filter theory of speech, which laid the foundation for the theory of acoustic speech production, forming the basis for modeling VTL and speaker characteristics prediction from acoustic cues.	Correlation Studies
Fitch and Giedd (1999)	Reported a strong correlation between speaker height and VTL (r = 0.93), founding the direction of height estimation using VTL.	Correlation Studies
González (2004)	Found correlations between formants and height or weight were generally weak, with stronger results in females than males ($r < 0.60$).	Correlation Studies
Hatano et al. (2012)	Showed that basic and intermediate features such as F0 and formants yield low correlation values in height prediction, especially vowel /e/.	Correlation Studies
Lammert and Narayanan (2015)	Demonstrated that the lowest RMSE values were associated with higher formants, which were less influenced by articulation.	Correlation Studies
Pisanski et al. (2014)	Summarized that F0 accounted for a maximum of 2% of the variance in human height, while formants explained up to 10% of height variation within sexes.	Correlation Studies
Kalluri et al. (2019)	Achieved a minimum RMSE of 6.1 cm using a combination of acoustic features.	SOTA Performance
Poorjam et al. (2015)	Achieved MAEs of 5.8 cm in female using ANNs and LSSVR, with a 60-dimensional vector as input.	SOTA Performance
Rajaa et al. (2021)	Achieved RMSEs of 6.0 cm for single-task model.	SOTA Performance
Story (2004)	Demonstrated that vowel distinctions are shaped by vocal tract configurations that mainly influence the F1-F3, higher formants (F4 and F5) enhance voice projection and resonance.	Research Gap Analysis
Dusan (2005)	Showed that higher feature dimensionality was linked to stronger correlations and captured more variability in speaker height.	Research Gap Analysis
Ganchev et al. (2010)	Demonstrated MFCCs' strong relevance in height prediction, with half of the top 50 features derived from them, while F0 also ranked top-10 3 times.	Research Gap Analysis

3 Methodology

This section outlines the overall methodology used to address the research question and evaluate the hypothesis. Subsection 3.1 introduces the TIMIT dataset used for training and testing, along with the rationale for its selection. Subsection 3.2 presents the phone-based approach, outlining its rationale, objectives, and the motivations for selecting this method. Subsection 3.3 presents the acoustic feature sets supported by a pilot study and justification for each choice, and then the regression models. Subsection 3.4 details the evaluation metrics and statistical tests employed to assess the significance of performance differences. Finally, Subsection 3.5 discusses the ethical considerations relevant to this study.

3.1 Dataset Description

The primary dataset used in this study is the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993). It is characterized by its well-controlled design, aimed at supporting both acoustic-phonetic research and the development and evaluation of ASR system. It features approximately five hours of high-quality recordings from 630 speakers across eight major American English dialect regions. Each speaker reads ten carefully selected, phonetically rich sentences, ensuring a comprehensive phonetic coverage and a balanced distribution of speech content. The dataset includes both male and female speakers and provides time-aligned transcriptions along with its high quality recording, making it suitable for task like vowel-based feature extraction and detailed acoustic analysis.

A key advantage of TIMIT is the inclusion of speaker metadata, such as age, gender, dialect region, education level, and height, which is particularly relevant for this study. Height information is often missing from other publicly available speech datasets, making TIMIT a rare and valuable resource for research exploring the relationship between acoustic features and physical speaker characteristics. Moreover, TIMIT's widespread use in related work (Dusan, 2005; Ganchev et al., 2010; Pellom & Hansen, 1997; Poorjam et al., 2015; Rajaa et al., 2021) provides a solid benchmark for comparison and justifies its selection for this study.

Each speaker folder of TIMIT contains 10 utterances, each utterance accompanied by an audio file (.WAV), a phoneme label file (.PHN), a word label file (.WRD) and a transcription file(.TXT). However, it does not contain annotation in Praat file format (.TextGrid), and the time-aligned phoneme annotations in phoneme label files use sample index instead of time (second), which require further data preprocessing before extracting features using Praat software.

3.2 Phone-based Approach

The phone-based approach was first proposed by Lamel and Gauvain (1995) and later adopted by Dusan (2005). Its central idea is to isolate non-linguistic, speaker-specific acoustic features by modeling individual phones rather than longer, semantically rich speech segments. Because phones operate at the sub-lexical level, this approach ensures that the input does not contain any semantic information. Instead, it captures only biometric traits, reducing the risk of inadvertently collecting conversational content. Therefore, this method allows the development and evaluation of height

prediction models that rely on minimal, linguistically impoverished input while retaining relevant acoustic markers.

In this study, the preset vowel subset provided by the TIMIT dataset are used as phone-based input for several reasons. First, this choice follows the precedent established by Lamel and Gauvain (1995) and Dusan (2005), ensuring methodological continuity and comparability. Second, TIMIT vowel subset includes a diverse set of vowel types. Compared to González (2004) and Hatano et al. (2012) who used only basic monophthongs /a/, /e/, /i/, /o/, and /u/, TIMIT vowel subset enable richer exploration of how different phonetic classes contribute to speaker height estimation by including also diphthongs. Finally, vowels are preferred over consonants because they naturally contain critical acoustic features (e.g., F0, formants) that are essential for experiments in this study. Consonants, by contrast, often lack these continuous resonant properties, making them less informative for this application. Therefore, throughout this work, the term phone refers exclusively to vowel segments rather than consonants or other sounds, as only vowels were included in the analysis due to their rich acoustic properties relevant for height prediction.

3.3 Feature Selection and Regression Models

3.3.1 Pilot Study of Feature Extraction and Software Selection

Due to the absence of Praat annotation files, I initially conducted a pilot study using utterances named SA1 and SA2, the universal utterances shared by all speakers, to explore whether basic features (F0, formants F1-F4) could be extracted directly from audio and phoneme label files. The aim was to use Python library librosa and Linear Predictive Coding (LPC), a widely used technique for formant estimation over the past decades (Rabiner, 1978) to extract F0 and F1-F4. If successful, this approach could provide a streamlined alternative to Praat for extracting basic acoustic features, eliminating the need for additional data preprocessing steps and script development typically required when using Praat.

However, the results presented several limitations. First, the F0 extraction using librosa's function was generally unreliable, as it occasionally produced missing or erratic values. More critically, the formant values estimated were often unstable and inconsistent that they did not align with the knowledge about formant values and vowels. In many cases, the F1-F3 values did not align with well-established acoustic patterns of vowel formants. For example, the expected F1-F3 ranges for vowel /i/ in American English were 280Hz, 2250Hz, and 2890Hz respectively (Ladefoged & Johnson, 2006), but some extracted values often ranged outside this range, such as close to 310Hz, 530Hz, and 1980Hz respectively. Furthermore, some vowels and F4 were often not estimated. Thus, these issues suggested that while this Praat-free approach was promising in concept, it was not a feasible replacement for Praat in reliable acoustic feature extraction.

3.3.2 Feature Selection

I selected fundamental frequency (F0) as a basic 1-dimension feature, formant frequencies (F1-F4) as intermediate-dimension features, and 13 Mel-frequency cepstral coefficients (MFCCs) to represent

high-dimension features, as these are commonly used in the literature, which is summarized in the table below:

Literature	F0	Formants	MFCCs
Dusan (2005)	✓	✓(F1-F4)	✓(10 dimensions)
Ganchev et al. (2010)	✓		✓
Hatano et al. (2012)	✓	✓(F1-F4)	
González (2004)	✓	✓(F1-F4)	
Lammert and Narayanan (2015)		✓(F1-F4)	

Table 3: Acoustic Features Used in Related Literature

Fundamental frequency (F0): Fundamental frequency (F0) is determined by the rate of vocal fold vibration and represents the lowest frequency of a periodic waveform. It is often equated with pitch, as the two typically correspond closely (Ladefoged & Johnson, 2006). F0 can be estimated using the following formula:

$$f_0 = \frac{1}{2L}\sqrt{\frac{T}{\mu}}$$

Formants: Resonant frequencies, also known as formants, correspond to overtones shaped by the vocal tract configuration and are critical in differentiating vowel qualities (Ladefoged & Johnson, 2006):

First Formant (F1): F1 is the lowest formant and is inversely related to vowel height. Vowels with a high tongue position, such as /i/ and /u/, have a low F1, while vowels with a low tongue position have a high F1. This pattern reflects articulatory vowel height. F1 can be perceived in creaky voice or simulated by tapping gently on the throat near the jaw while maintaining a vowel posture.

Second Formant (F2): F2 is associated with the frontness or backness of the tongue. Front vowels, for example, /i/ and /e/, have a high F2, while back vowels, such as /u/ and /o/, have a low F2. F2 is more perceptible when vowels are whispered, due to the absence of vocal fold vibration.

Third Formant (F3): F3 contributes to finer distinctions in vowel quality, though it plays a less prominent role than F1 and F2 in basic vowel identification. It is more difficult to isolate perceptually, but it adds nuance to the overall vowel sound.

Fourth Formant (F4): While F4 does not play a major role in distinguishing vowel quality, it may reflect speaker-specific traits such as head or vocal tract size. F4 is suggested to be a normalization reference for other formants (F1-F3).

Mel-Frequency Cepstral Coefficients (MFCCs): The computation of Mel-Frequency Cepstral Coefficients (MFCCs) involves the following steps:

1. **Framing and Windowing:**
$$x_w[n] = x[n] \cdot w[n]$$

2. **Fourier Transform (FFT):**
$$X[k] = \sum_{n=0}^{N-1} x_w[n] \cdot e^{-j2\pi kn/N}$$

3. **Power Spectrum:**
$$P[k] = |X[k]|^2$$

4. **Mel Filter Bank Application:** Apply a set of triangular filters spaced according to the Mel scale to obtain the Mel energy coefficients $M[m]$.

5. **Logarithm of Mel Energies:**
$$\log M[m]$$

6. **Discrete Cosine Transform (DCT):**
$$c_n = \sum_{m=1}^{M} \log M[m] \cdot \cos\left[\frac{\pi n}{M}(m-0.5)\right], \quad n = 1,2,...,N_{\text{MFCC}}$$

Here, c_n is the n^{th} MFCC coefficient, M is the number of Mel filters, and N_{MFCC} is the number of coefficients extracted. In this study, I chose to extract 13 MFCCs because this number strikes a practical balance between capturing essential spectral information and maintaining computational efficiency. The lower-order coefficients effectively represent the overall spectral shape and the separation of source (vocal folds) and filter (vocal tract) characteristics, which are crucial for modelling speech. Additionally, extracting 13 coefficients is a well-established convention in speech processing, commonly used in both classical and modern systems, making it a reliable and comparable choice for analysis and modelling.

3.3.3 Regression Models

Acoustic Feature Set	Simple Linear	Multiple Linear	Random Forest
F0	✓		
Formants (F1-F4)	✓	✓	✓
MFCCs (1-13)	✓	✓	✓

Table 4: Overview of Regression Models Applied to Each Acoustic Feature Set

This study employs both linear and non-linear regression models. Simple linear regression is applied to evaluate each acoustic feature individually. By using only one feature at a time, this approach offers a clear and interpretable demonstration of predictive power for each variable in isolation. While this simplicity helps highlight distinct patterns across features, it is less capable of capturing

the full relationship between feature sets and speaker height, and therefore is expected to yield lower predictive performance overall. Multiple linear regression is included because the seed paper by Dusan (2005) used it to demonstrate the correlation between speaker's height and multiple acoustic features. In contrast, non-linear regression is incorporated to reflect the methodological direction of Ganchev et al. (2010), who used more advanced non-linear approaches. This study applies random forest regression as a non-linear method that can model complex, non-linear dependencies without requiring extensive parameter tuning, providing a more interpretable alternative to techniques such as support vector machines or deep neural networks.

Simple Linear Regression: Simple linear regression is employed to model the relationship between a single acoustic feature and speaker height. This model assumes a linear dependency of the form, where x is each individual feature:

$$\hat{y} = \beta_0 + \beta_1 x$$

Multiple Linear Regression: Multiple linear regression is used to evaluate how a set of acoustic features can jointly predict speaker height. The model assumes a linear combination of inputs, where x are the selected set of acoustic features:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Random Forest Regression: Random forest regression is applied to capture potential non-linear relationships between acoustic features and height. As an ensemble method, it builds multiple decision trees during training and outputs the average of their predictions, improving robustness and reducing overfitting. Unlike linear models, it can model complex interactions and non-additive effects among features. Additionally, random forest regression provides feature importance scores, which help identify which and how acoustic features contribute most to height prediction.

3.4 Evaluation Methodology

Root Mean Squared Error (RMSE): RMSE is a widely used metric for evaluating the accuracy of regression models and is particularly common in assessing the performance of SOTA height estimation systems. It quantifies the average magnitude of prediction errors by taking the square root of the mean of the squared differences between predicted and actual values. A lower RMSE value indicates better predictive performance, with smaller errors between predicted and actual speaker heights. Mathematically, RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

To validate the results, both the Wilcoxon Signed-Rank Test and the Friedman Test are employed. These non-parametric statistical tests are used to compare paired or grouped observations—Wilcoxon for pairwise comparisons and Friedman for comparisons across three or more re-

lated conditions, such as different models or feature sets. The specific use of each test is detailed in Subsection 1.3.

3.5 Ethics and Research Integrity

This research was conducted in alignment with the ethical standards set by the faculty ethical guidelines. No ethical approval was required as the study used TIMIT, a publicly available datasets that do not include personally identifiable information. All procedures align with the institutional policies on data handling and responsible research practices.

To promote transparency and collaboration, all scripts and models developed in this study are shared via [https://github.com/stellasiu/2025thesisdemo], accompanied by documentation and usage instructions. The repository includes licensing and citation guidelines. All experiments were reproducible using standard Python libraries and Praat, which is an open-source tool.

However, the potential biases and fairness limitations of the TIMIT dataset, as well as the responsibilities involved in reproducing or replicating this study, will be discussed in detail, given that height prediction from phone-based input still relies on sensitive biometric data extracted from speech signal.

3.5.1 Bias and Fairness

Although the TIMIT dataset is widely used in speaker profiling research, it presents several potential biases and fairness limitations.

First, the dataset is linguistically biased, as it is composed exclusively of American English speakers. This limits its representativeness across English varieties and other languages, potentially reducing the generalizability of the findings. Since this study does not include any cross-lingual datasets or experiments to evaluate performance in other linguistic contexts, which constitutes a clear limitation and suggests an important direction for future research.

Furthermore, the dataset is demographically imbalanced, containing only about 30% female speakers and an unequal distribution of speakers across dialect regions. Although this imbalance is not explicitly accounted for in the experimental design as gender and dialect region are not distinguished in the analysis, it is reflected in documented disparities in prediction performance reported in SOTA studies, including substantial differences between male and female speakers. For example, Rajaa et al. (2021) reported a notable difference in RMSE between male speakers (8.1 cm) and female speakers (6.0 cm), highlighting the impact of biased dataset on model accuracy. To mitigate this limitation, future work could explore data augmentation strategies, such as synthetically increasing the representation of female speakers by sampling from the normal distribution of human height, to create a more balanced dataset and better evaluate model performance across demographic groups.

3.5.2 Responsibility of Reproduction and Replication

When reproducing or replicating this study on acoustic feature-based height prediction, it is critical to acknowledge the responsibility inherent in handling sensitive data. The main challenges concern both biometric privacy and conversational privacy.

Acoustic features inherently encode biometric information that can distinguish personal speech characteristics, making it possible to infer or identify individuals based on distinctive combinations of features. Beyond height, these features can be used for broader speaker profiling in forensic phonetics field to estimate attributes such as age, gender, and even psychological states (Leemann, Perkins, Buker, & Foulkes, 2024). In fact, SOTA artificial intelligence methods have demonstrated the potential to reconstruct a speaker's facial appearance by learning associations between vocal and facial features (Leemann et al., 2024), further increasing the sensitivity and potential misuse of such data. Therefore, the potential social impact of biometric privacy breaches is substantial if such data is to be leaked or misused. Consent is also a critical concern when collecting new data, and participants must be fully informed about how their recordings will be stored, processed, and potentially shared.

Conversational privacy relates to the semantic content embedded in speech recordings. Some datasets, such as TIMIT, mitigate this risk by requiring participants to read aloud pre-defined prompts with rich phonetic coverage. However, datasets containing spontaneous or conversational speech may include highly sensitive personal details—such as identification numbers, financial information, or private narratives—that should not be collected without explicit, informed consent. In practical implementations of speaker profiling, live speech data often contains such content. This risk can be minimized by collecting phone-based utterances directly, rather than extracting segments from broader conversational recordings. A further ethical issue involves the risk of misuse for surveillance. If these technologies are deployed without clear safeguards, they could be used to monitor individuals based on incidental speech content, including humour, political expression, or controversial opinions. Such practices may undermine freedom of expression and erode trust in biometric technologies. The phone-based approach adopted in this study provides one way to mitigate these risks: because phones are sub-lexical units, they do not contain any interpretable semantic content, substantially reducing the likelihood of such misuse and better protecting conversational privacy.

In short, this study suggests that phone-based modelling offers a promising approach to mitigating conversational privacy risks in acoustic feature-height prediction research. Nonetheless, biometric privacy concerns remain unavoidable, underscoring the importance of careful handling, secure storage, and transparent governance of such data.

4 Experimental Setup

In this chapter, I will a comprehensive overview of the experimental setup throughout this research. In subsection 4.1, I will introduce the preliminary setup, including data splitting, data preprocessing steps, feature extraction pipeline, the data cleaning steps in subsection 4.1. The following subsections (4.2, 4.3, and 4.4 are organized by acoustic feature groups, each describing the experimental setup, including implementation details, execution time per model, memory usage, and the total number of features used. For Random Forest regression, relevant hyperparameters are also reported to ensure reproducibility.

Subsection	Acoustic Features	Regression Models
4.2	F0	Simple Linear Regression
4.3	Formants (F1-F4)	Simple Linear, Multiple Linear, and Random Forest Regression
4.4	MFCCs(1-13)	Simple Linear, Multiple Linear, and Random Forest Regression

Table 5: Overview of Experiment Setup

The self-written Praat script used for feature extraction, as well as the Python scripts for phoneme alignment, data cleaning and regression models in this section, can be found in this repository: Demo (hereafter referred to as "Demo"). These resources are provided to support full reproducibility and to facilitate further analysis or adaptation by other researchers, particularly given the limited availability of publicly shared Praat scripts for separately extracting features.

4.1 Data Preparation

4.1.1 Data Splitting

The TIMIT dataset provides a predefined train/test split, consisting of 462 speakers in the training set and 168 speakers in the test set. Both sets include coverage of all eight dialect regions (DR1-DR8) and contain a balanced mix of male and female speakers. Within each set, speakers are organized into dialect region folders and labeled according to gender and speaker ID.

I adopted this predefined split for the experiments for three main reasons. First, it effortlessly ensures representation across all dialect regions and speaker demographics. Second, it maintains a practical train/test ratio of approximately 73.3% to 26.7%, which supported an effective and straightforward hold-out validation process. Third, as noted by Garofolo et al. (1993), this split guaranteed that there was no overlap in sentence text between the training and test sets, ensuring that model evaluation was not biased by repeated content.

However, since dialect regions were not a critical factor in the current experiments, the files were reorganized into simplified train and test directories locally. For clarity and consistency, all files were renamed using the format shown in the following table (see timitrename.py in Demo):

Division	Explanation
Set	Train or Test
Dialect Region	Dialect regions 1-8
Speaker ID	Provided by TIMIT dataset, including the gender prefix
File Part	File name of the audio file, e.g. SA1, SX403

Table 6: Overview of Renaming Pattern

4.1.2 Phoneme Alignment

To extract acoustic features using Praat, it is first necessary to align the phonetic transcriptions from the TIMIT dataset with their corresponding audio recordings. Each audio file in the dataset is accompanied by a .PHN file, which lists phone-level segmentations with corresponding start and end times. However, these timestamps are expressed in sampling rate units rather than seconds, and no annotated alignment files in .TextGrid format are provided.

Therefore, to prepare the data for phonetic alignment and analysis, I developed a multi-stage preprocessing pipeline using custom Python scripts (see sr2s.py, phn2txt.py and txt2grid.py in Demo). The first step involved converting all start and end times in the .PHN files from sample indices to seconds by dividing each value by TIMIT's sampling rate (16,000). This conversion was implemented in a batch process to efficiently handle all files. Next, I designed another Python script to systematically convert the modified .PHN files into .txt files with a simplified and more readable format, maintaining the structure of phone labels along with their corresponding time intervals in seconds. Then, in order to facilitate compatibility with Praat, the .txt files were programmatically converted into .TextGrid format. This final step involved generating properly structured .TextGrid files for each audio sample, using the same file names as the corresponding recordings, and preserving all phone boundary information within time-aligned tiers.

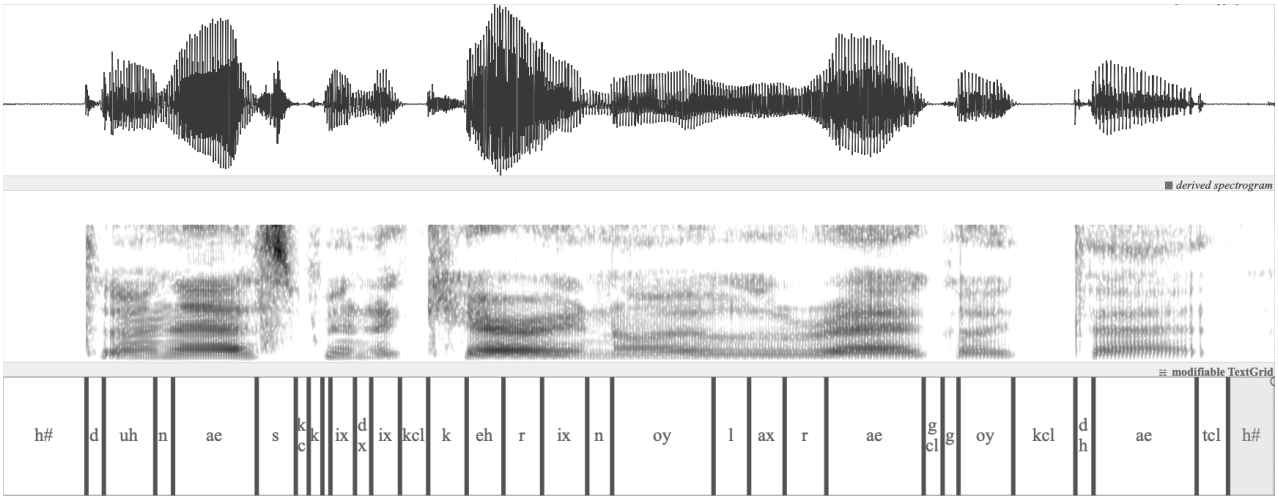


Figure 1: A Spectrogram with Transcription of One of the Sentence-texts All Speakers Spoke

4.1.3 Height Conversion

In addition to preparing the phoneme annotations, the height metadata from the TIMIT dataset was also processed. Since only speaker height was required for the analysis and the values were originally recorded in inches, a conversion to centimeters (cm) was performed for all speakers, rounded to two decimal places (see heightconversion.py in Demo). The converted data were saved to a separate .csv file to facilitate the later data cleaning process. This conversion was necessary because the RMSE used in the evaluation is measured in cm, and aligning the units ensured consistency in model output and interpretation.

4.1.4 Feature Extraction

I extracted the feature sets using different software tools, depending on the type of acoustic feature. I used Praat to extract the fundamental frequency (F0) and formants (F1-F4). Then I used the Python library librosa to compute the Mel-Frequency Cepstral Coefficients (MFCCs).

- Fundemantal Frequency (F0): I extracted F0 using a self-written Praat script (see f0_all.praat in Demo). This script automated the extraction of mean F0 values from phones (vowel segments) in multiple .TextGrid and .WAV file pairs within a specified directory, and saved the results to individual .csv files. It began by defining the input configuration and generating a list of files to process. A key feature of the script was its use of a predefined, space-separated list of ARPAbet vowel labels from the TIMIT dataset, which ensured that only intervals labeled as vowels were analyzed. For each vowel segment, the script computed the mean F0 within a typical pitch range of 75-600 Hz, helping to ensure more reliable measurements.

- Formants (F1-F4): I extracted formant frequencies (F1-F4) using a Praat script that I modified from one originally developed by Joey Stanley available online (see formants_all.praat in Demo). Similar to the F0 extraction script, this script automated the extraction of formant values from TIMIT vowel segments by processing paired .TextGrid and .WAV files within a specified directory. For each vowel segment, the script retrieved the start and end times, calculated the duration, and measured the mean frequencies of F1, F2, F3, and F4 over that interval. The results were saved in .csv format for each file to support further analysis and model integration. This approach improved upon the pilot study by ensuring consistent extraction of all four formant values, thereby guaranteeing feature completeness.

- MFCCs: I extracted MFCCs using a custom Python script (see mfcc_phoneme.py in Demo), built with the librosa library. This script processed .WAV files and used accompanying .TextGrid annotations to extract MFCC features from vowel-labeled phoneme segments. I used the default librosa settings, with a hop length of 512 and 13 MFCCs per segment. For each vowel segment, the script identified the corresponding time interval, computed its duration, and extracted MFCCs using a standard short-time Fourier transform (STFT) window. It then calculated the mean value for each MFCC dimension across the segment and saved the results to a .csv file for each input, supporting further analysis and modeling.

4.1.5 Data Cleaning Procedures

After extracting all the acoustic features, the resulting values were saved as individual .csv files named according to the audio file naming pattern described in subsection 4.1.1. To prepare the data for experiment, these files needed to be merged into a single .csv file per data splitting set per feature set (train/test), so that each feature type had a corresponding train.csv and test.csv file that included speaker height in cm. For each feature set, I performed this data cleaning and merging using a custom Python script (see praat_sm_clean.py in Demo), which parsed the filenames based on the renaming pattern (see Table 6) and merged the height metadata accordingly.

4.2 Experiment 1: F0

For the first experiment, I conducted a simple linear regression using only the F0 feature set to examine its standalone predictive power and to provide a baseline for comparing against more complex feature sets. F0, which represents vocal pitch, has been widely studied and reported to have a moderate correlation with speaker height (r = 0.59) (Dusan, 2005). While not the most informative feature, analyzing F0 in isolation helps contextualize the contribution of additional acoustic features and supports evaluation of how feature dimensionality affects prediction performance.

It was modelled using the most basic regression technique: simple linear regression. The implementation was carried out in Python 3.11 using the scikit-learn library and scipy library (see f0_sr.py in Demo). The mean F0 values extracted were used as the only predictor of speaker height. The evaluation was based on one key metric: RMSE. The model was trained and tested on the predefined TIMIT train/test split, with no additional feature engineering or hyperparameter tuning applied. The script also generated .csv outputs for statistical testing and a heatmap of RMSE values sorted from lowest to highest for easier interpretation.

The experiment was executed on a MacBook Air with an Apple M3 chip and 24 GB of RAM. This setup required minimal runtime (under one minute per run).

4.3 Experiment 2: Formants

The second experiment followed a similar structure and included three sub-experiments using formant frequencies as predictors of speaker height (see formants_sr.py, formants_mr.py, and formants_rf.py in Demo). All experiments were conducted on the same CPU-based machine setup (MacBook Air with an Apple M3 chip and 24 GB RAM). The runtimes were approximately one to two minutes per run.

First, I applied simple linear regression using each of the four formants (F1-F4) individually as predictors. Next, I employed multiple linear regression to model all four formants together as multivariate input. Finally, I used random forest regression to capture potential non-linear relationships using all four formants. The scripts saved outputs to .csv files for statistical analysis and generated RMSE heatmaps for each regression model. In the case of simple linear regression, the heatmap displayed RMSE values for each of the four formants across all 20 phones, and then for each phone, the script computed mean RMSE values of F1-F4 for statistical analysis. For the multiple linear

regression, a single RMSE heatmap was generated across all 20 phones using the four formants as multivariate input.

The random forest regression was implemented using scikit-learn version 1.3.2. The dataset structure remained consistent across experiments as in linear regression models. I used a fixed random state of 42, 20 estimators, and a maximum depth of 10. A separate heatmap displayed RMSEs for each phone-formant combination. An additional plot visualized feature importance scores, highlighting the relative contribution of each formant to height prediction per phone. Random forest also computed mean RMSE values across phones for statistical analysis.

4.4 Experiment 3: MFCCs

For the third experiment (see mfcc_sr.py, mfcc_mr.py, and mfcc_rf.py in Demo), it also included three sub-experiments, with the key difference being the use of 13 MFCCs instead of formants as predictors of speaker height. All experiments were conducted on the same CPU-based machine setup (MacBook Air with an Apple M3 chip and 24 GB RAM). The runtimes were approximately one to two minutes per run.

First, I performed simple linear regression using each of the 13 MFCC coefficients independently as predictors of speaker height. This was followed by multiple linear regression, where all 13 MFCCs were simultaneously used as multivariate input to capture joint effects. Lastly, I applied random forest regression to explore potential non-linear relationships using the complete MFCC feature set.

Each model's output was saved as a .csv file for subsequent statistical analysis, and corresponding RMSE heatmaps were generated to visualize performance. For simple linear regression, the heatmap illustrated RMSE scores across all 13 MFCCs for each of the 20 phones. Mean RMSE values across MFCCs were then computed per phone to facilitate comparative statistical testing. In the multiple regression case, a single heatmap visualized RMSEs across all phones using the multivariate MFCC input.

Random forest regression was implemented using scikit-learn (v1.3.2), maintaining consistency in data structure with the previous experiment. The model was configured with a fixed random seed of 42, 20 estimators, and a maximum tree depth of 10. A separate heatmap visualized RMSEs for each phone–MFCC pair, while an additional plot showed feature importance scores, indicating the relative contribution of each MFCC to height prediction by phone. Mean RMSE values were also computed for use in statistical validation.

5 Results

In this chapter, detailed results are presented for all three acoustic feature sets: F0 (5.1.1), formants (5.1.2), and MFCCs (5.1.3). Subsection 5.1 first reports the performance of the regression models, including tables summarizing the maximum and minimum RMSE values and their corresponding phones for each feature or feature sets, then organized the results by acoustic features, in the order of F0, formants and MFCCs. Finally, Subsection 5.2 presents the statistical test results, organized by the hypotheses outlined in Subsection 1.3. It includes summary tables of p-values and integrates the findings to evaluate all three hypotheses. Full visualizations, including RMSE heatmaps and feature importance scores heatmaps, are provided in the Appendix.

5.1 Performance

Feature	Max. RMSE	Phone	Min. RMSE	Phone
F0	10.52	/ax-h/	7.2	/ux/
F1	9.28	/ax-h/	7.14	/aw/
F2	9.44	/ax-h/	8.13	/aw/
F3	9.37	/ax-h/	8.08	/uw/
F4	9.10	/ax-h/	7.61	/uw/
MFCC1	9.60	/ax-h/	8.51	/aw/
MFCC2	9.56	/ax-h/	7.95	/ay/
MFCC3	9.59	/ax-h/	8.47	/aw/
MFCC4	9.58	/ax-h/	8.44	/aw/
MFCC5	9.64	/ax-h/	7.80	/ux/
MFCC6	9.67	/ax-h/	8.43	/oy/
MFCC7	9.59	/ax-h/	8.26	/ow/
MFCC8	9.30	/ax-h/ & /uh/	8.17	/ae/
MFCC9	9.60	/ax-h/	8.37	/ux/
MFCC10	9.68	/ax-h/	8.01	/aw/
MFCC11	9.69	/ax-h/	8.16	/aw/
MFCC12	9.62	/uh/	8.01	/oy/
MFCC13	9.59	/ax-h/	8.07	/ux/

Table 7: Overview of Simple Linear Regression Results

Examining Table 7, F0 shows the highest maximum RMSE at 10.52 cm and a low minimum RMSE at 7.20 cm, indicating the unstable performance across phones. The average maximum RMSE increases with the complexity of acoustic features: 9.30 cm for formants, and 9.59 cm for MFCCs. In terms of minimum RMSE, formants performed best at 7.74 cm, followed by MFCCs at 8.2 cm.

These results suggested that increasing feature complexity did not lead to improved prediction performance under simple linear regression. Interesting, assessing the minimum RMSE values, F1, F4, MFCC2, and MFCC5 each fall below 8 cm. Notably, these features also exhibit relatively high importance scores in the random forest regression model for height prediction. This suggests that they may be particularly effective for phone-based height estimation compared to other features.

/ax-h/ consistently yields the highest RMSE across nearly all features, indicating it has the least predictive power and showing a stable trend of poor performance. In contrast, the phones associated with the minimum RMSE such as /ux/, /uw/, and /aw/, vary considerably across features, suggesting that high predictive power is not consistently tied to specific phones in simple linear regression. This variability highlights an instability in which phone yields the best predictions, unlike the consistent underperformance observed with /ax-h/.

Feature	Max. RMSE	Phone	Min. RMSE	Phone
Formants	8.63	/ax-h/	7.19	/aw/
MFCCs	9.23	/ax-h/	6.88	/aw/

Table 8: Overview of Multiple Linear Regression Results

Examining Table 8, /ax-h/ consistently corresponds to the maximum RMSE, suggesting it is the most difficult phone for predicting height, even when using multiple features. Likewise, /aw/ consistently corresponds to the minimum RMSE, indicating strong and stable predictive performance across all feature sets.

The maximum RMSE increases with the complexity of acoustic features, even when using a multiple linear regression model. However, all three feature sets show significant improved performance compared to their results under simple linear regression. Notably, MFCCs achieve the lowest minimum RMSE of 6.88 cm among all linear regression models.

Feature	Max. RMSE	Phone	Min. RMSE	Phone
Formants	8.93	/ax-h/	7.45	/ae/
MFCCs	9.16	/ax-h/	7.18	/aw/

Table 9: Overview of Random Forest Regression Results

Examining Table 9, /ax-h/ consistently corresponds to the maximum RMSE, suggesting its difficulty in height prediction even when using a non-linear regression model. All feature sets, except MFCCs, show worse performance compared to the multiple linear regression model, suggesting that non-linear regression model does not enhance prediction accuracy universally.

5.1.1 F0 Results

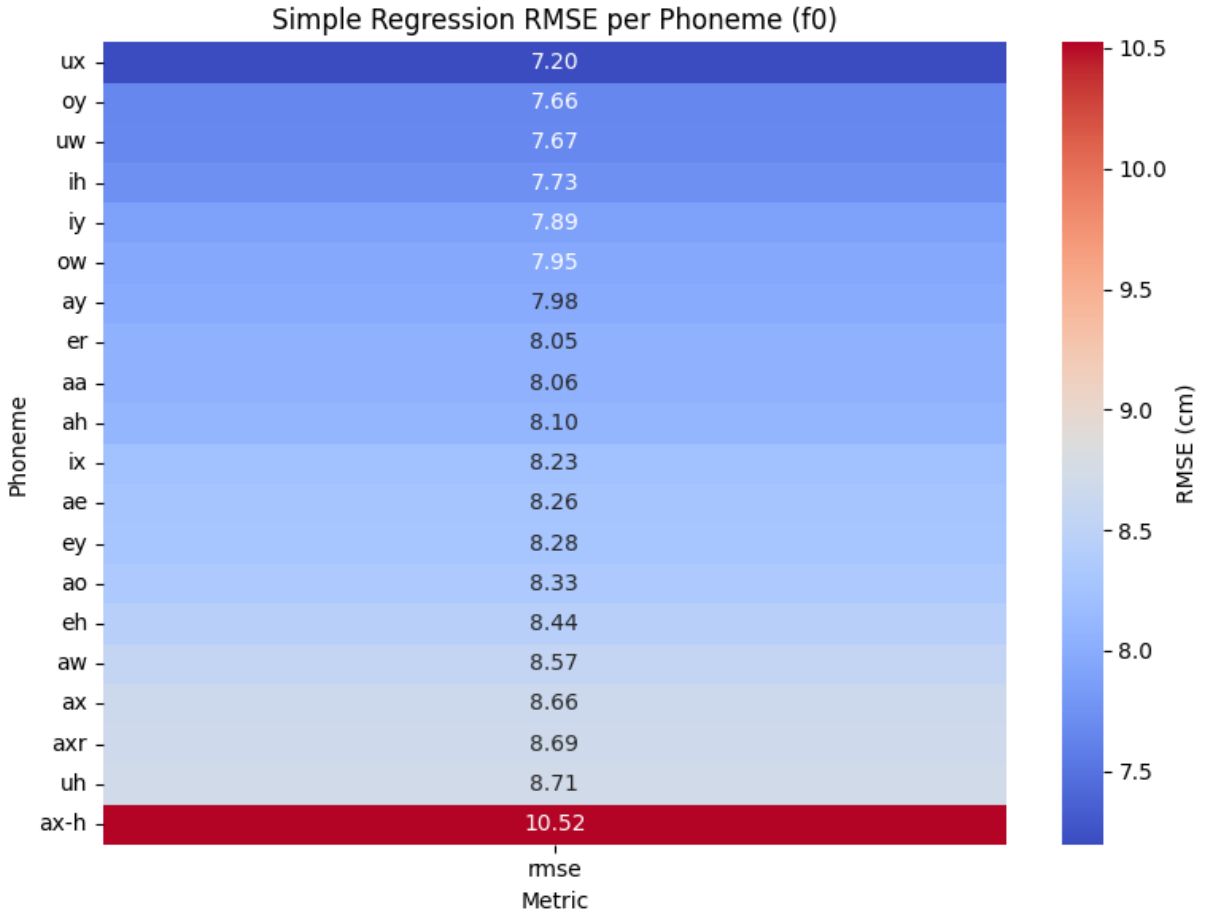
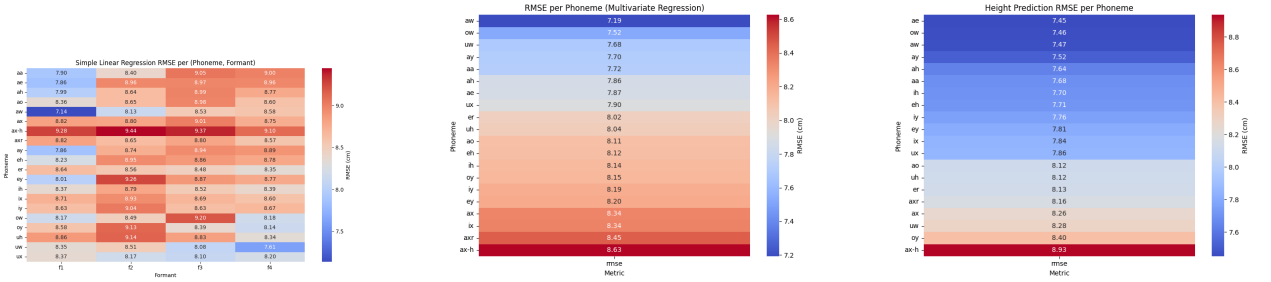


Figure 2: F0 RMSE per Phone for Height Prediction Using Simple Linear Regression Model

In Figure 7, the five phones with the lowest RMSE values are /ux/, /oy/, /uw/, /ih/, and /iy/ respectively, while the five phones with the highest RMSE values are /ax-h/, /uh/, /axr/, /ax/, and /aw/. These show that close front vowels perform better than open back vowels. Notably, /ax-h/ exhibits a significantly higher RMSE: 1.81 cm greater than /uh/, which has the second highest RMSE value. Excluding /ax-h/, the differences in RMSE among the remaining phones are all less than 1 cm, indicating relatively similar performance across those cases.

Close rounded vowels (/ux/, /oy/, /uw/) demonstrate the lowest prediction errors, and close front unrounded vowels (/ih/, /iy/) also demonstrate low prediction errors, while /ax-h/ shows significantly poorer performance, indicating that vowel articulation position influences height prediction accuracy when using fundamental frequency alone. However, F0 has the highest maximum RMSE at 10.52 and the second lowest minimum RMSE at 7.2 cm, the wide range of RMSE demonstrates the unstable predictive power in the most basic acoustic feature set.

5.1.2 Formant Results



(a) Formants RMSE using Simple Linear Regression

(b) Formants RMSE using Multiple Linear Regression

(c) Formants RMSE using Random Forest Regression

Figure 3: RMSE Heatmaps of Formants across 20 phones

Figure 4: Formants Feature Importance per Phone per Feature

Full-size versions of the plots are available in A.2.

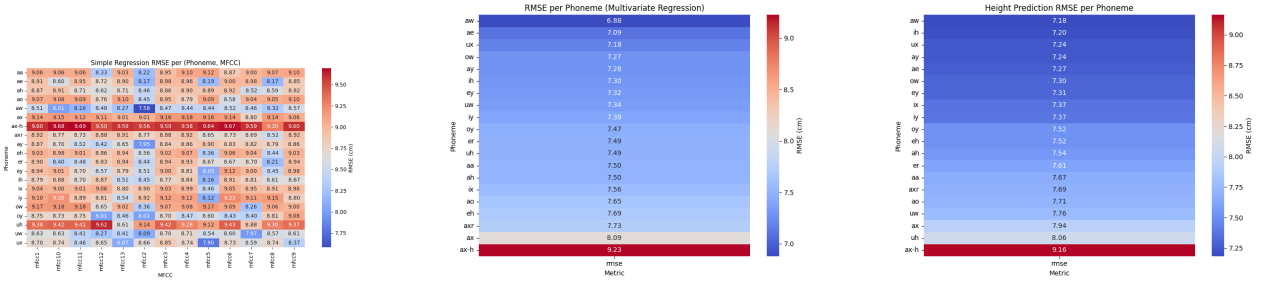
Simple Linear Regression: F1 achieves the best performance among formants with five phones showing RMSE below 8 cm, particularly for open vowels. The phone /ax-h/ consistently shows the worst performance across all formants. These best performed phones, /aa/, /ae/, /ah/, /ao/, and /aw/, are all open back vowels. This finding contrasts with the results for F0. Conversely, the five phones with the highest RMSE values for F4 are /ux/, /uw/, /uh/, /oy/, and /ow/, which are all back rounded vowels. The result particularly reveals an inverse pattern: phones that performed well using F1 tended to perform poorly using F4, and vice versa. These results suggest that F1 and F4 capture distinct and potentially complementary articulatory cues relevant to height estimation.

Multiple Linear Regression: The phone /aw/ achieves the lowest RMSE while /ax-h/ remains the poorest performer. RMSE differences among phones are minimal, suggesting comparable performance across most vowels when combining all formants. Unlike the simple regression results, no clear phone-specific pattern emerged in the multivariate formant analysis.

Random Forest Regression: Performance is distributed across five ranges with /ae/ showing best results and /ax-h/ worst. Notable gaps exist between /oy/ and /ax-h/, though no clear phonetic patterns emerge. Feature Importance Score show that F1 and F4 demonstrate high importance scores while F2 and F3 show minimal contribution. An inverse relationship exists between F1 and F4 importance, suggesting complementary roles in the model.

Phonetic Interpretation: The consistently strong performance of open vowels such as /aa/, /ae/, and /aw/ may be linked to their open vocal tract configuration, which facilitates a more stable and unobstructed airflow. This openness may enhance the clarity and consistency of formant frequencies, especially F1, making them more reliable indicators of vocal tract length, and thus, speaker height. In contrast, centralized and reduced vowels like /ax-h/ involve shorter, more variable articulatory gestures and less distinct resonance patterns, which likely leads to degraded predictive performance. However, attributing this to specific articulatory mechanisms remains speculative, as the methodological focus on averaged features fails to account for any theoretical interpretations. Furthermore, the observed inverse relationship between F1 and F4 effectiveness may reflect a trade-off between the articulatory space, which influences F1, and laryngeal/posterior resonances, which influences F4, underscoring how vowel articulation shapes the acoustic cues available for height estimation.

5.1.3 MFCCs Results



(a) MFCCs RMSE using Simple Linear Regression

(b) MFCCs RMSE using Multiple Linear Regression

(c) MFCCs RMSE using Random Forest Regression

Figure 5: RMSE Heatmaps of MFCCs across 20 phones

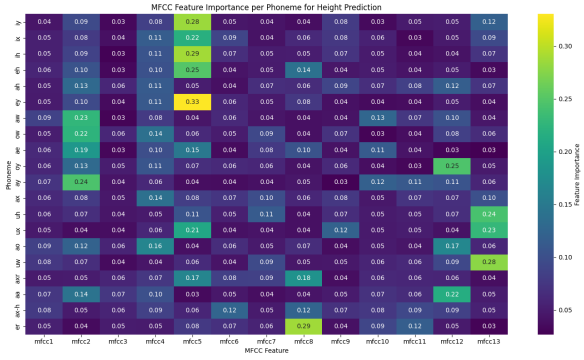


Figure 6: Formants Feature Importance per Phone per Feature

Full-size versions of the plots are available in A.3.

Simple Linear Regression: MFCC2 and MFCC5 generally outperform other features in height prediction while MFCC1 and MFCC3 exhibited poor performance. The phone /aw/ achieved the best performance across all MFCCs, followed by /oy/ as the second best. Conversely, /ax-h/ performed the worst across all MFCC features, with /uh/ as the second worst.

Multiple Linear Regression: Except for /ax/ and /ax-h/, all other phones achieved RMSE values below 8 cm, with /aw/ recording the lowest minimum RMSE at 6.88 cm. MFCCs indicated strong overall performance when compared to other feature sets using the multiple linear regression model.

Random Forest Regression: /ax-h/ and /aw/ respectively had the maximum and minimum RMSE. While RMSE of /ax-h/ slightly dropped by 0.07 cm compared to multiple linear regression model, RMSE of /aw/ increased by 0.3 cm. Same as multiple linear regression model, 18 phones achieved RMSE values below 8 cm, indicating generally good performance compared to other features and regression models. The feature importance scores indicate that MFCC2 and MFCC5 contribute more significantly than other MFCCs, consistent with the performance results from simple linear regression.

Phonetic Interpretation: The phone /aw/ consistently achieves the lowest RMSE, even with MFCCs. This may reflect its long duration, dynamic articulatory movement, and rich spectral content as a diphthong vowel, though further empirical testing is needed to confirm this. The transition from a low back to a high front rounded position results in substantial changes in the spectral envelope may be well captured by MFCCs, especially the mid-order coefficients that encode such spectral dynamics. Stable and information-rich patterns may make phones, such as /aw/, highly discriminative for speaker profiling tasks, though this hypothesis requires acoustic validation. Conversely, /ax-h/ performs poorly across all models, including those using MFCCs. As a reduced and centralized vowel, /ax-h/ tends to be short, acoustically weak, and variable across speakers and contexts. These traits may lead to noisy or unstable MFCC representations, and do not reliably reflect speaker-specific anatomy. This may explain the persistent high RMSE associated with /ax-h/, even in more complex models.

Unlike formants, MFCCs are more robust across all phones because they summarize spectral energy over the entire frequency range. This holistic view makes them less sensitive to articulatory differences between vowels and more capable of capturing speaker-intrinsic characteristics regardless of vowel type. As a result, MFCC-based models demonstrate relatively stable performance across diverse phonemes, including diphthongs, high vowels, and even reduced vowels.

MFCC2 and MFCC5, which yield best performance in simple linear regression and show significant contribution in random forest regression, capture broad spectral envelope patterns in the lower frequency range, which corresponds to articulatory features like vowel height and frontness that are indirectly related to VTL. Since these dimensions are more stable and less affected by coarticulation than higher-order coefficients, MFCC2 and MFCC5 provide stronger, more consistent cues for speaker characteristics like height.

5.2 Statistical Test Results

Hypothesis	Test	IV	DV	p-value	Statistically Significant
H1	Friedman	F0, Formants, MFCCs (from SR)	(Mean) RMSEs of phones	0.0000012	✓
	Wilcoxon	Formants, MFCCs (from MR)	RMSEs of phones	0.0001335	✓
H2	Friedman	20 Phones	RMSE of SR	0.0000012	✓
	Wilcoxon	20 Phones	RMSEs of MR	0.0001335	✓
	Wilcoxon	20 Phones	RMSEs of RF	0.0000362	✓
H3	Wilcoxon	SR, MR	Mean RMSEs of Formants	0.0000019	✓
	Wilcoxon	SR, MR	Mean RMSEs of MFCCs	0.0000019	✓
	Wilcoxon	Formants, MFCCs	ΔRMSE	0.0000038	✓

Table 10: Overview of Statistical Results

Full-size versions of the plots are available in B.

The statistical tests summarized in Table 10 confirm that all observed differences are statistically significant across hypotheses H1 to H3 as all of them yield p-values well below the standard threshold ($p < 0.05$). The validation of hypotheses is presented below.

5.2.1 Validation of Hypothesis 1

H1: High-dimensional acoustic features (MFCCs) will produce lower RMSE values than basic (F0) and intermediate features (formants) across most phones when using linear regression models.

Validation criteria: It is considered validated if statistically significant results ($p < 0.05$) allow analysis of the conditions under which each feature set predicts height most effectively.

As this hypothesis is based on Dusan (2005), who reported that the correlation between speaker height and acoustic features increases with the dimensionality of those features using multiple linear regression model, the correlation showed in the findings is list below:

Feature	Pearson's r
MFCC(1-10)	0.7426
Formants (F1-F5)	0.7264
F0	0.5880

Table 11: Overview of Dusan (2005)'s Correlations

This hypothesis is accepted since it is statistically significant. As shown in the boxplots, the maximum RMSE values generally fall outside the upper whisker—except for formants in the multiple linear regression model—indicating that these values are outliers relative to the rest of the data and can be treated accordingly.

The simple linear regression results appear contradictory and challenge the findings of Dusan (2005) and align more closely with Ganchev et al. (2010), who ranked F0 highly in feature relevance. However, although F0 exhibited the lowest median RMSE in simple linear regression, its high maximum value (10.52 cm) and wide range indicate unstable performance across phones, calling into question whether it can be considered the best-performing feature overall. On the other hand, MFCCs clearly outperformed formants in multiple linear regression, supporting the idea that higher-dimensional features more effectively capture height-related information. This suggests that more complex regression models may be better suited to uncover meaningful correlations in the data.

Cross-referencing with the heatmaps (A) identifies /ax-h/ as the phone consistently associated with the highest RMSE values. Its repeated outlier status across all feature sets and regression models indicates extremely poor predictive reliability. This may stem from its acoustic variability, reduced articulation as a centralized vowel), or limited height-discriminative cues. These results highlight the importance of vowel identity and phonetic structure in shaping model performance.

5.2.2 Validation of Hypothesis 2

H2: The phone /ax-h/ will consistently exhibit the highest RMSE across all feature sets (Dusan, 2005), due to its status as a reduced vowel characterized by high articulatory and acoustic variability. Open vowels (/aa/, /ae/, /aw/) are expected to yield lower RMSE values, as they involve greater vocal tract expansion compared to close vowels.

Validation criteria: /ax-h/ consistently yields higher RMSEs, and at least one open vowel (/aa/, /ae/, /aw/) consistently yields lower RMSEs across all three tests.

As illustrated in the line plots, /ax-h/ not only consistently produces the highest RMSE, but also exhibits an unusual trend: it is the only vowel where MFCCs yield higher RMSEs than formants across all regression models. Thus, the hypothesis regarding /ax-h/ is supported. Regarding open vowels, /aw/ and /aa/ demonstrate stable and strong performance, particularly in multiple linear and random forest regression. While /ux/ shows the lowest RMSE in the F0-based simple regression, /aw/ consistently performs well across all three models. Therefore, /aw/ meets the criterion of yielding consistently low RMSE values, supporting the second part of the hypothesis. In summary, H2 is

accepted based on consistent trends across all evaluations.

Indeed, other diphthongs, such as /ay/ and /ow/, also demonstrate strong performance in more complex models. Their extended duration and broad articulatory movement likely enhance the clarity of formant transitions and the richness of the spectral envelope, reinforcing their link to speaker-specific anatomical characteristics and leading to better prediction performance.

5.2.3 Validation of Hypothesis 3

H3: Multiple linear regression will outperform simple linear regression, but the improvement will be more pronounced for high-dimensional features (MFCCs with 13 dimensions) than for intermediate features (formants with four dimensions) (Dusan, 2005).

Validation criteria: MR outperform SR in both tests, and improvement of MFCCs is more pronounced than of formants ($\Delta\text{RMSE} = \text{RMSE}_{\text{MultipleRegression}} - \text{RMSE}_{\text{SimpleRegression}}$).

As shown in the boxplots, multiple linear regression consistently yields lower RMSEs than simple linear regression, validating the superiority of the more complex model. Furthermore, the reduction in RMSE for MFCCs is significantly greater than that observed for formants except for one outlier in ΔRMSE. These results underscore the role of both feature dimensionality and model complexity in improving the accuracy of speaker height estimation.

In particular, these findings emphasize that richer acoustic representations benefit more from sophisticated modelling techniques, such as multiple regression, which can capture complex interdependencies between features. This supports the idea that both the choice of features and the regression method used are critical for maximizing predictive performance in speech-based biometric applications.

In short, the performance evaluation across regression models reveals that feature complexity and model selection play critical roles in speaker height prediction. In the next chapter, these findings will be discussed in detail and examined in relation to research questions.

6 Discussion

After analysing the results presented in Section 5, it becomes clear that the complexity of acoustic features significantly influences height prediction accuracy at the phone level across linear regression models. These findings directly address our primary research question: How does acoustic feature complexity affect height prediction accuracy when comparing basic features (F0), intermediate features (formants), and high-dimensional features (MFCCs) across different regression models at the phone-based level?

Overall, while simple linear regression favoured F0 with unstable performance, high-dimensional features such as MFCCs demonstrated superior performance under more complex models like multiple linear regression and random forest regression. Multiple linear regression demonstrates better ability to capture complementary spectral information than random forest regression when random forest regression configuration is not optimized. This indicates that the benefits of feature complexity are best realized when paired with models capable of capturing multivariate relationships, and suggests that linear regression may be more effective than non-optimized non-linear approaches in this context.

The following discussion interprets these results in relation to the sub-research questions outlined in 1.3, with particular consideration of using phones as minimal input for height prediction and reference to critical literature reviewed in 2. It also reflects on several limitations and then concludes with suggestions for improving the generalizability and interpretability of acoustic-based height estimation.

6.1 Impact of Feature Complexity

How does the use of different feature sets (basic, intermediate, high-dimensional) impact RMSE across phones and regression models? In particular, do high-dimensional features (MFCCs) consistently outperform simpler features?

Although H1 is accepted, the simple linear regression results appear to contradict with Dusan (2005). While F0 showed the lowest RMSE (7.20 cm), and the F0 of some phones yielded notably low RMSE values compared to the mean RMSE of formants and MFCCs in simple linear regression, its high maximum value (10.52 cm) and wide range likely reflect that, although it is a single, clear feature linked to vocal fold vibration and speaker height—making it easy for a simple model to fit—it also varies greatly across speech contexts. This variability makes F0 unstable and not an optimal predictor at the phone level overall. This demonstrates that the simplest regression model combined with the most basic feature set is not practical for accurately predicting height at the phone-based level. Moreover, the mean RMSE values of intermediate and high-dimensional feature sets were worse in simple linear regression overall, indicating that regardless of which feature set is chosen, simple linear regression is not an effective approach for height prediction.

When focusing specifically on multiple linear regression, the results are in fact aligned with Dusan (2005) who used multiple linear regression as primary methodology, as high-dimensional features demonstrated better prediction accuracy. This suggests that the correlation between acoustic

features and height can translate into predictive performance, but only under appropriate modelling conditions.

On the other hand, although random forest regression is a non-linear model, its results do not align closely with the findings of Ganchev et al. (2010), who identified F0 as a top-ranked acoustic feature. Instead, simple linear regression showed performance patterns more consistent with that ranking. This indicates that predictive outcomes are not solely determined by feature relevance but are also influenced by the regression method applied.

Formants performed particularly well for open vowels like /aw/ and /ae/, and their interpretability allows specific features such as F1 or F4 to serve as effective minimal input. This makes formants a potential compromise between simplicity and accuracy. MFCCs using multiple linear regression achieved a minimum RMSE falling below the 7 cm RMSE value mark typically reported by SOTA systems noted in Section 2. Therefore, this result should be considered outstanding given that it is achieved without advanced machine learning architectures. This highlights phone-based MFCCs combined with multiple linear regression as a strong and practical minimal-input approach for height prediction.

6.2 Phone-specific Patterns

Are there specific phones for which height can be predicted most or least accurately, and do these patterns align with articulatory openness (open vs. closed vowels) and phonetic reduction?

The consistent underperformance of /ax-h/ across all feature sets and models, which may reflect its articulatory feature as a reduced, centralized vowel being produced with minimal articulatory effort and often appears in unstressed syllables. This may also reflect the shorter duration, weaker intensity, and highly variable acoustic realizations of /ax-h/. These properties potentially obscure speaker-specific traits and make height estimation from /ax-h/ particularly unreliable. However, additional acoustic analysis is necessary to substantiate this claim as current study only captures mean values and cannot directly verify the proposed articulatory properties.

/aw/ demonstrated good predictive power. It may reflect dynamic articulatory properties from a low back to a high front position and a shift from unrounded to rounded articulation, though this requires further acoustic validation. Other diphthongs, such as /ay/ and /ow/, also yield a better performance in more complex models, but the performance is rather unstable. Lee, Potamianos, and Narayanan (2014) stated that diphthongs are notably characterized by dynamic formant transitions, but the rate of change varies across different diphthongs and serves as a key cue for distinguishing them since the onset and offset portions of diphthongs do not consistently align with the monophthongs typically used to transcribe them phonetically depending on speaker and context.

Therefore, on one hand, this different onset-offset portions of diphthong may be able to include wide, clear, and unique overall spectral shape, strengthening its association with speaker-specific anatomical features. On the other hand, the inherent variability and complexity of diphthongs may lead to distinctive performance patterns for each diphthong individually. Interestingly, this finding partially contrasts with the correlation results reported by Dusan (2005), where the highest phone-

based correlation between speaker height and MFCC features was found for /iy/ ($r = 0.7254$), a closed front unrounded vowel. This discrepancy suggests that strong feature-height correlations at the phone level do not always translate into high predictive accuracy in regression models, particularly among non-reduced vowels. Furthermore, while /aw/ consistently achieved the lowest RMSE values, the current analysis cannot definitively attribute this to specific articulatory properties because the extraction of mean acoustic values precludes direct measurement of the dynamic spectral changes characteristic of diphthongs.

Taken together, these findings demonstrate that the performance of individual phone varies. This highlights the importance of considering vowel identity and articulatory structure when selecting input units for speaker attribute estimation, and also the need to further examine whether temporal acoustic dynamics correlate with prediction accuracy to validate these hypothesized mechanisms when aiming for minimal input. Overall, /aw/ emerges as the most effective candidate. In contrast, /ax-h/ should be excluded from the input selection due to low predictive value.

6.3 Regression Method Insights

How does the performance (RMSE) of linear regression models (simple and multiple) compare to that of a non-linear model (random forest regression) in predicting speaker height from acoustic features across phones?

Multiple linear regression clearly outperforms simple linear regression, as confirmed by the validation of H3. The improvement, expressed as ΔRMSE, is especially pronounced for high-dimensional features such as MFCCs compared to intermediate features like formants. Simple linear regression's inability to account for feature interactions or assign differential weights across multiple predictors and underutilize the richness of complex features, leading to suboptimal RMSE performance. This is solely reflected on the phone /aw/. Although the diphthong /aw/ consistently shows strong predictive performance in more complex models, its performance under simple linear regression is comparatively less impressive. This can be attributed to the limitations of simple regression in capturing the spectral features inherent in diphthongs. Simple linear regression, which models each acoustic feature independently, is suggested to fail to account for this temporal complexity and the interactions among multiple features.

When comparing multiple linear regression to random forest regression, the results show that multiple linear regression consistently yields better predictive accuracy. The greater improvement of MFCCs with multiple linear regression may reflect their ability to capture complementary spectral information that linear combinations can explore, as MFCCs encode a broader and more nuanced representation of the spectral envelope, though this requires further investigation.

MFCCs exhibit less stability compared to formants, as reflected by their lower minimum but higher maximum mean RMSE values. While MFCCs may achieve better performance under certain conditions, their variability suggests they may be less reliable across all phones. In contrast, formants appear to offer more consistent performance in non-linear models such as random forest regression. Notably, feature importance scores from the random forest model clearly capture the empirically observed complementary relationship between F1 and F4, supporting prior findings and

highlighting formants may serve as a relatively stable and interpretable input for height prediction. Nonetheless, further investigation is needed to clarify and validate these patterns.

Although multiple linear regression appears to outperform non-linear regression in this study, no definitive conclusions can be drawn regarding the overall effectiveness of linear versus non-linear approaches because the random forest configuration used in this study represents a conservative implementation. For instance, parameters such as random state, number of estimators, and maximum depth were not systematically tuned, which may have limited its performance. The superior performance of multiple linear regression may reflect either genuine suitability for this task or simply suboptimal random forest parameters.

6.4 Physiological Insights from Feature Weights

Which acoustic features contribute most significantly to height prediction in the random forest regression model, and how do feature importance patterns relate to the acoustic-physiological mechanisms?

For formants, an inverse relationship between F1 and F4 importance is observed in the random forest regression, with further support from the simple linear regression results: phones that perform well using F1 tend to perform poorly with F4, and vice versa. This pattern aligns with the findings of Lammert and Narayanan (2015) and Barreda (2016) that F4 can act as a complementary cue to F1—often inversely—for height prediction. Furthermore, the findings of González (2004) found that the F2 of /e/ (corresponding to /eh/ in this study and the TIMIT dataset) is strongly correlated with speaker height across sexes. However, both the RMSE values and feature importance scores for /eh/ across regression models in this study do not support that claim.

For MFCCs, MFCC2 and MFCC5 show comparatively higher importance scores in the random forest regression model, indicating a stronger contribution to height prediction. MFCC2 appears to contribute more to open back vowels, while MFCC5 shows greater relevance for close front vowels, suggesting that different MFCC components may encode vowel-specific spectral cues linked to speaker height. However, unlike formants, simple linear regression does not clearly reflect this pattern, as performance varies across phones without highlighting these specific coefficients.

6.5 Limitations

The first limitation concerns methodology. Although both core reference studies—Dusan (2005) and Ganchev et al. (2010)—also use the TIMIT dataset, making this study comparable, it is important to note that TIMIT is based on American English. As a result, the findings may not generalize to other datasets, particularly those in different languages, due to phonological differences.

The second limitation involves the scope of feature extraction. This study focuses only on a limited set of acoustic features. Although this set includes the most commonly chosen features, they may not capture all speaker-specific characteristics. The absence of more nuanced or data-driven representations may have constrained the predictive ceiling of the models used.

The third limitation lies in the rigidity of the statistical framework. Although predefined validation criteria helped maintain analytical consistency, they may have limited flexibility in interpreting the results. Future research could adopt a more exploratory perspective to uncover the conditions in which various acoustic features are most effective.

The fourth limitation includes the conservative implementation of random forest regression configuration. As the model was not extensively tuned or tested across multiple configurations, conclusions regarding the relative performance of linear versus non-linear approaches remain limited. The apparent superiority of multiple linear regression over random forest regression may reflect suboptimal random forest parameters rather than a fundamental limitation of non-linear methods.

In short, the results show that high-dimensional features like MFCCs achieve the best height prediction when used with multiple linear regression. /aw/ offer the most reliable input while /axh/ perform poorly. Multiple linear regression consistently outperforms simple models, and feature importance patterns align with known acoustic-physiological cues. These findings highlight the value of combining appropriate features, models, and phone types for accurate and interpretable height estimation.

7 Conclusion

This thesis investigated the research question **"How does acoustic feature complexity affect height prediction accuracy when comparing basic features (F0), intermediate features (formants), and high-dimensional features (MFCCs) across different regression models at phone-based level?"**. In this conclusion, I will summarize the main contributions, discuss future research directions, and reflect on the broader impact of this work.

7.1 Summary of the Main Contributions

This study revealed how acoustic feature complexity, regression model choice, phone-specific characteristics, and individual acoustic features collectively influence speaker height prediction at the phone level.

First, this study examined the interaction between feature complexity and modelling approach in relation to height-feature correlation. While F0 produced the lowest RMSE in simple linear regression, it was outperformed by MFCCs in multiple linear regression. This supports the view that high-dimensional features deliver superior performance when used with models capable of capturing multivariate relationships. Notably, MFCCs combined with multiple linear regression achieved an RMSE below 7 cm—comparable to SOTA without relying on advanced machine learning architectures.

Second, it investigated phone-specific reliability in height prediction. /aw/ was identified as the most reliable phone for height prediction due to its dynamic articulation and rich spectral cues, while /ax-h/ consistently showed the poorest performance across all features and models. This confirms H2 and highlights the importance of articulatory openness and vowel reduction in predictive accuracy.

Third, it provides insights on regression comparison. Multiple linear regression consistently outperformed simple linear regression, but limited to prove also outperform random forest regression. Improvements were especially notable for high-dimensional features, validating H3. This suggests that the feature-target relationship may be linear, non-linear relationship is recommended to be further explored.

Fourth, it explored the physiological interpretability of features. Feature importance analysis in random forest regression revealed an inverse relationship between F1 and F4, in line with prior findings, and identified MFCC2 and MFCC5 as key contributors depending on vowel type.

In short, this study contributes to the enhancement of minimal-input height prediction systems. It demonstrates that carefully selected acoustic features (e.g. MFCCs), combined with informative phones (e.g. /aw/) and appropriate modelling (e.g. multiple linear regression), can achieve prediction accuracy comparable to complex architectures used by SOTA. This positions phone-based acoustic modelling as a viable, interpretable, and conversational privacy-preserving approach to height estimation.

7.2 Future Work

While the findings of this study offer valuable insights into acoustic-based height prediction, several limitations must be acknowledged to contextualize the results as in 6.5. Nevertheless, these limitations also suggest several directions for future work:

First, there is a need for cross-linguistic or cross-dataset validation. Future research should extend the current study to datasets in other languages or beyond the TIMIT corpus to evaluate the generalizability of the findings. Phonological differences across languages and dialects—such as vowel inventories and phonotactic constraints—may influence the stability and effectiveness of phone-based height cues identified in this study.

Second, future work could explore more dynamic and data-driven inputs, such as SOTA embeddings at the phone level. This would enable a direct comparison between traditional acoustic features and modern learned representations, potentially revealing richer patterns in speaker-height prediction.

Third, the current use of random forest regression was limited to a single configuration and was not extensively tuned. Further exploration of non-linear models with varying parameters, such as tree depth, estimators, and random seeds, is necessary to draw firmer conclusions about the comparative strengths of linear and non-linear modelling approaches in this context. Other non-linear regression models can also be employed to compare with random forest regression results.

Fourth, this study heavily relied on a hypothesis-driven framework with rigid pre-set validation criteria. Although it successfully contributes to the understand of speaker height estimation by affirming and providing evidence to some prior studies, the strict requirement may have constrained the opportunity for data-driven insights. Future research could adopt a more exploratory perspective to examine conditions of height prediction.

Last but not least, a correlation study between temporal acoustic dynamics and prediction accuracy is proposed to test the hypothesized mechanisms, especially the difference between monophthongs and diphthongs. This proposed analysis could further clarify whether specific acoustic changes over time are linked to predictive performance and strengthen the evidence for their role in height estimation.

7.3 Impact & Relevance

In short, this study contributes to the development of minimal-input height prediction systems by showing that height can be accurately estimated using only short, phone-level acoustic segments. This approach offers linguistically impoverished input that protects conversational privacy while still enabling reliable biometric analysis.

It further positions phone-based acoustic modelling as a viable, interpretable, and conversational privacy-conscious method for estimating speaker height, particularly valuable in contexts such as forensic phonetics or biometric authentication where full recordings may be limited or contain sensitive conversations. With further model development, even a brief snippet of speech, such as an

isolated /aw/ vowel produced by a suspect or victim, could be sufficient to estimate the speaker's height and link this information to existing biometric databases for investigative or identification purposes. Another potential application can be seen in countries such as the Netherlands, where passports include the holder's height as part of identity verification. Although the current model is not yet sufficiently robust for operational deployment in high-stakes scenarios, advancing this approach will eventually enable reliable use cases such as verifying the claimed identity of suspected undocumented immigrants by comparing predicted height against official records.

Furthermore, this study demonstrates that phone-based input can deliver predictive performance comparable to that of current industry frameworks, which predominantly rely on deep learning models trained on full speech signals. This suggests that phone-level acoustic segments has potential to replace full speech recordings in certain applications and be integrated with advanced machine learning techniques to develop efficient height estimation systems that conserve conversational privacy.

However, it is important to acknowledge that while conversational privacy is improved, biometric privacy concerns remain. This limitation is inherently difficult to avoid, as the acoustic features required for height estimation inherently capture aspects of vocal tract anatomy that also contribute to distinctive voice patterns, making individuals potentially identifiable.

References

Barreda, S. (2016). Investigating the use of formant frequencies in listener judgments of speaker size. *Journal of Phonetics*, *55*, 1–18.

Barreda, S., & Predeck, K. (2024). Inaccurate but predictable: Vocal-tract length estimation and gender stereotypes in height perception. *Journal of Phonetics*, *102*, 101290.

Despres, T., Constantino, M. A., Lizola, N. Z., Romero, G. S., He, S., Zhan, X., ... Bernd, J. (2024). " my best friend's husband sees and knows everything": A cross-contextual and cross-country approach to understanding smart home privacy. *Proceedings on Privacy Enhancing Technologies*.

Dusan, S. (2005). Estimation of speaker's height and vocal tract length from speech signal. In *Interspeech* (pp. 1989–1992).

Fant, G. (1960). Acoustic theory of speech production, s'-gravenhage. *Mouton and Co*.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *106*(3), 1511–1522.

Ganchev, T., Mporas, I., & Fakotakis, N. (2010). Audio features selection for automatic height estimation from speech. In *Artificial intelligence: Theories, models and applications: 6th hellenic conference on ai, setn 2010, athens, greece, may 4-7, 2010. proceedings 6* (pp. 81–90).

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Pallett, D. S., Dahlgren, N. L., Zue, V., & Fiscus, J. G. (1993). Timit acoustic-phonetic continuous speech corpus. *(No Title)*.

González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of phonetics*, *32*(2), 277–287.

Hatano, H., Kitamura, T., Takemoto, H., Mokhtari, P., Honda, K., & Masaki, S. (2012). Correlation between vocal tract length, body height, formant frequencies, and pitch frequency for the five japanese vowels uttered by fifteen male speakers. In *Interspeech* (pp. 402–405).

Hernandez-de Menendez, M., Morales-Menendez, R., Escobar, C. A., & Arinez, J. (2021). Biometric applications in education. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *15*, 365–380.

Kalluri, S. B., Vijayasenan, D., & Ganapathy, S. (2019). A deep neural network based end to end model for joint height and age estimation from short duration speech. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6580–6584).

Ladefoged, P., & Johnson, K. (2006). *A course in phonetics* (Vol. 3). Thomson Wadsworth Boston.

Lamel, L., & Gauvain, J.-L. (1995). A phone-based approach to non-linguistic speech feature identification. *Computer Speech & Language*, *9*(1).

Lammert, A. C., & Narayanan, S. S. (2015). On short-time estimation of vocal tract length from formant frequencies. *PloS one*, *10*(7), e0132193.

Lass, N. J., & Davis, M. (1976). An investigation of speaker height and weight identification. *The Journal of the Acoustical Society of America*, *60*(3), 700–703.

Lee, S., Potamianos, A., & Narayanan, S. (2014). Developmental acoustic study of american english diphthongs. *The Journal of the Acoustical Society of America*, *136*(4), 1880–1894.

Leemann, A., Perkins, R., Buker, G. S., & Foulkes, P. (2024). *An introduction to forensic phonetics and forensic linguistics*. Taylor & Francis.

Mohammed, S. M., & Ali, O. (2024). Human biometric identification: Application and evaluation. *IJECS*, *6*(2), 131–152.

Pellom, B. L., & Hansen, J. H. (1997). Voice analysis in adverse conditions: the centennial olympic park bombing 911 call. In *Proceedings of 40th midwest symposium on circuits and systems. dedicated to the memory of professor mac van valkenburg* (Vol. 2, pp. 873–876).

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J., Röder, S., Andrews, P. W., ... Feinberg, D. R. (2014). Vocal indicators of body size in men and women: a meta-analysis. *Animal Behaviour*, *95*, 89–99.

Poorjam, A. H., Bahari, M. H., Vasilakakis, V., & Hamme, H. V. (2015). Height estimation from speech signals using i-vectors and least-squares support vector regression. In *2015 38th international conference on telecommunications and signal processing (tsp)* (pp. 1–5).

Rabiner, L. R. (1978). *Digital processing of speech signals*. Pearson Education India.

Rajaa, S., Van Tung, P., & Siong, C. E. (2021). Learning speaker representation with semi-supervised learning approach for speaker profiling. *arXiv preprint arXiv:2110.13653*.

Rudrapal, D., Das, S., Debbarma, S., Kar, N., & Debbarma, N. (2012). Voice recognition and authentication as a proficient biometric tool and its application in online exam for ph people. *International Journal of Computer Applications*, *39*(12), 6–12.

Schilling, N., & Marsters, A. (2015). Unmasking identity: Speaker profiling for forensic linguistic purposes. *Annual Review of Applied Linguistics*, *35*, 195–214.

Story, B. H. (2004). Vowel acoustics for speaking and singing. *Acta Acustica united with Acustica*, *90*(4), 629–640.

Yee, K., & MacKown, P. (2009). Detecting and preventing cheating during exams. *Pedagogy, not Policing*, *141*.

Appendices

A RMSE & Feature Importance Score Heatmaps

A.1 F0

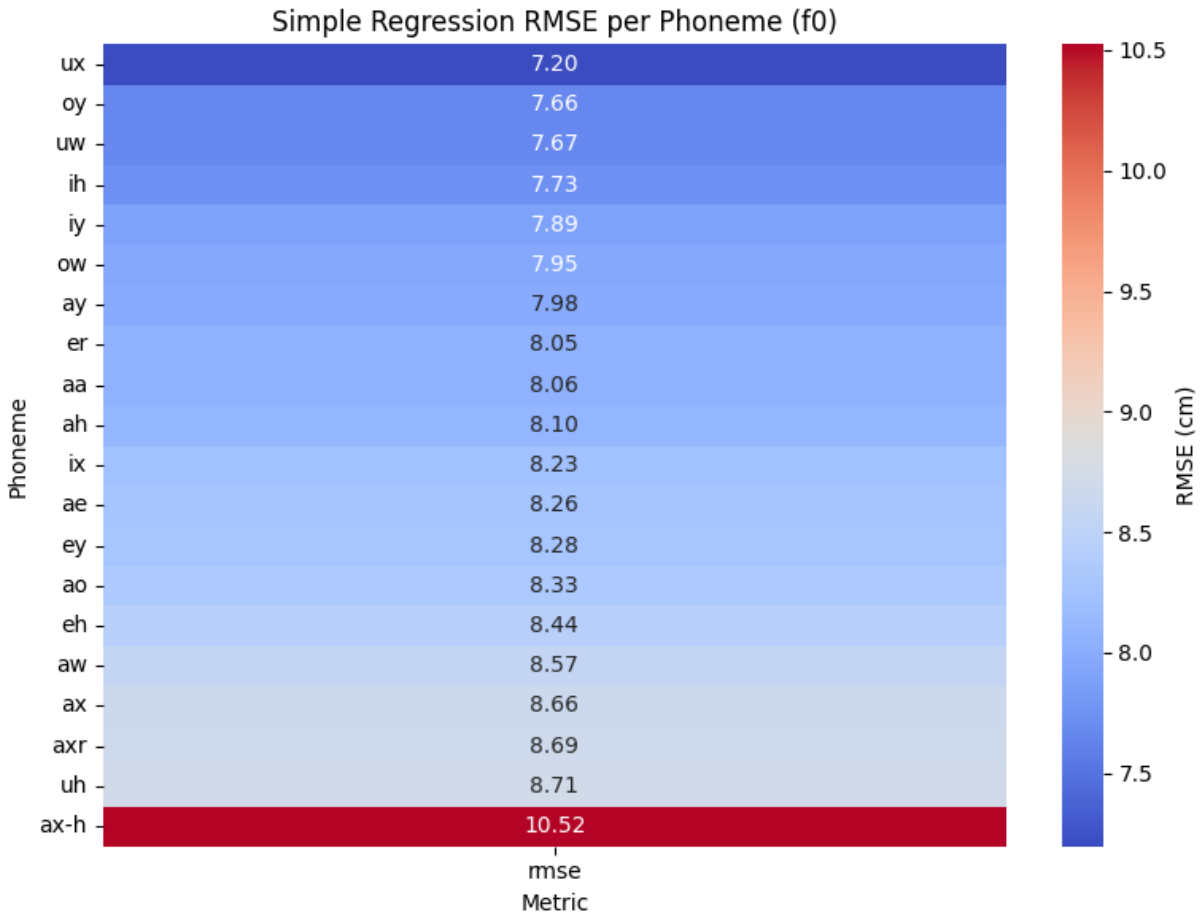


Figure 7: F0 RMSE per Phone for Height Prediction Using Simple Linear Regression Model

A.2 Formants

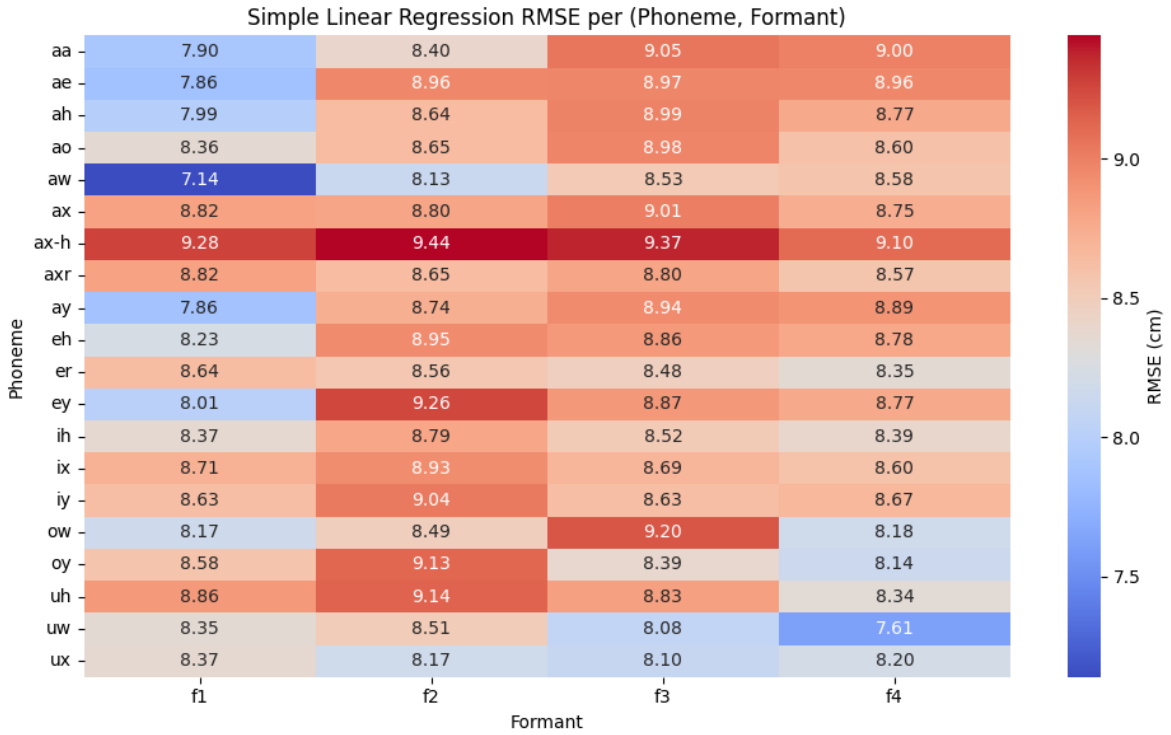


Figure 8: Formants RMSE per Phone per feature for Height Prediction Using Simple Linear Regression Model

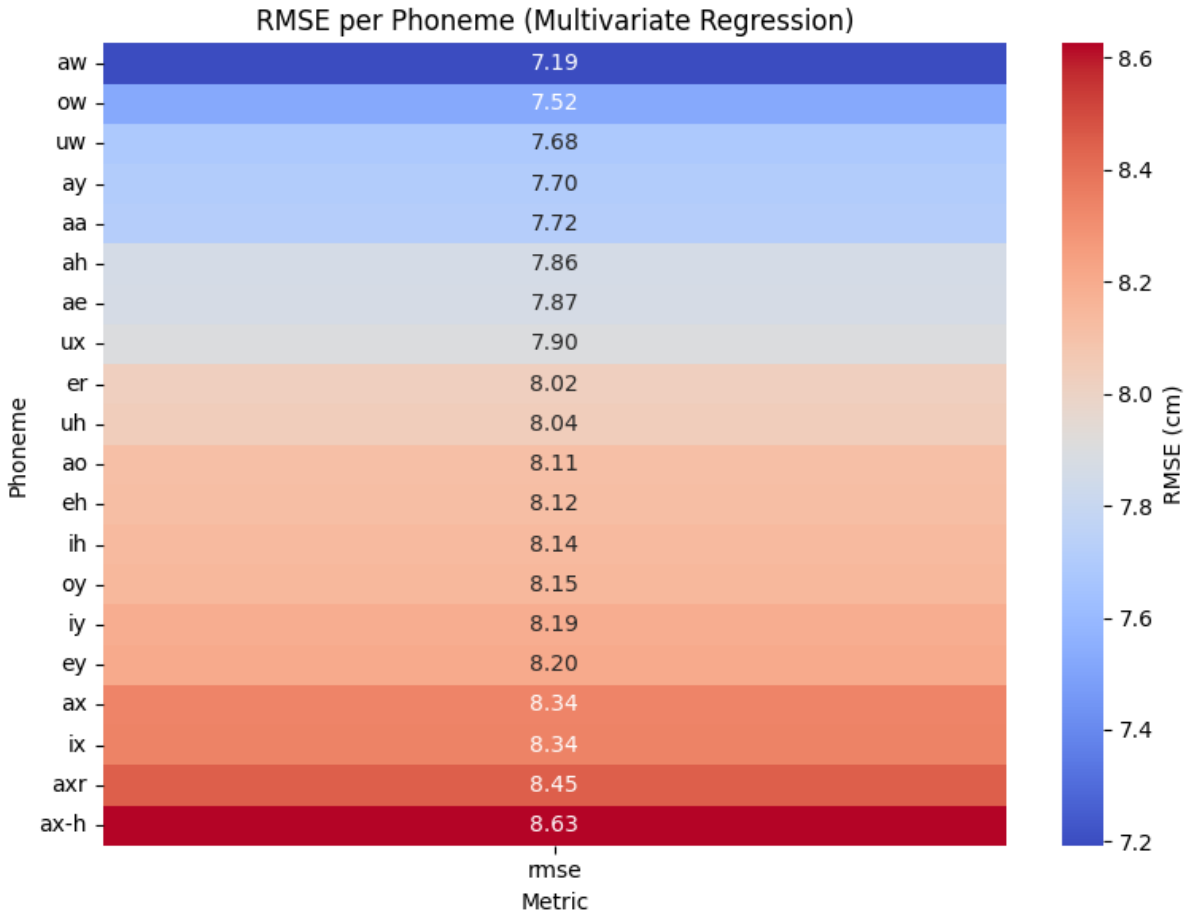


Figure 9: Formants RMSE per Phone for Height Prediction Using Multiple Linear Regression Model

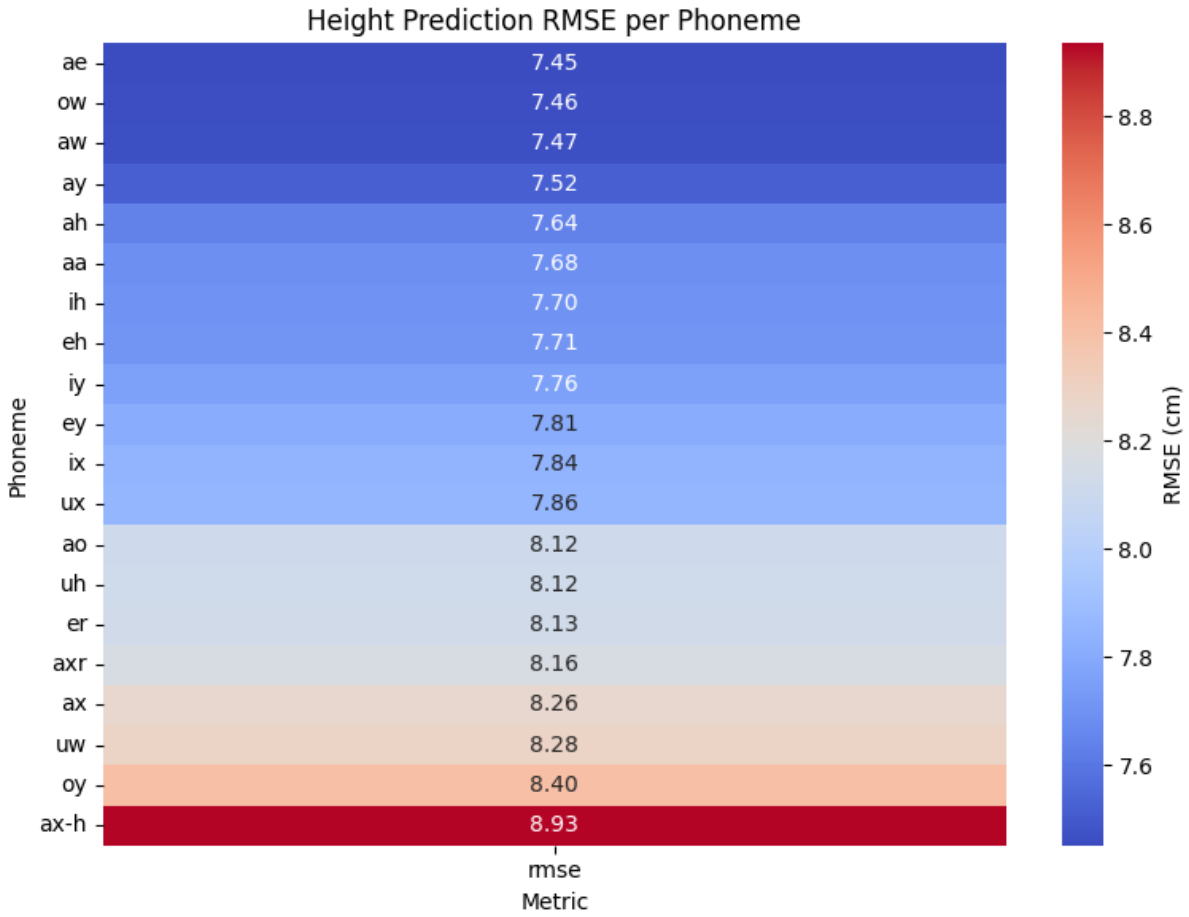


Figure 10: Formants RMSE per Phone for Height Prediction Using Random Forest Regression Model

Figure 11: Formants Feature Importance per Phone per Feature for Height Prediction Using Random Forest Regression Model

A.3 MFCCs

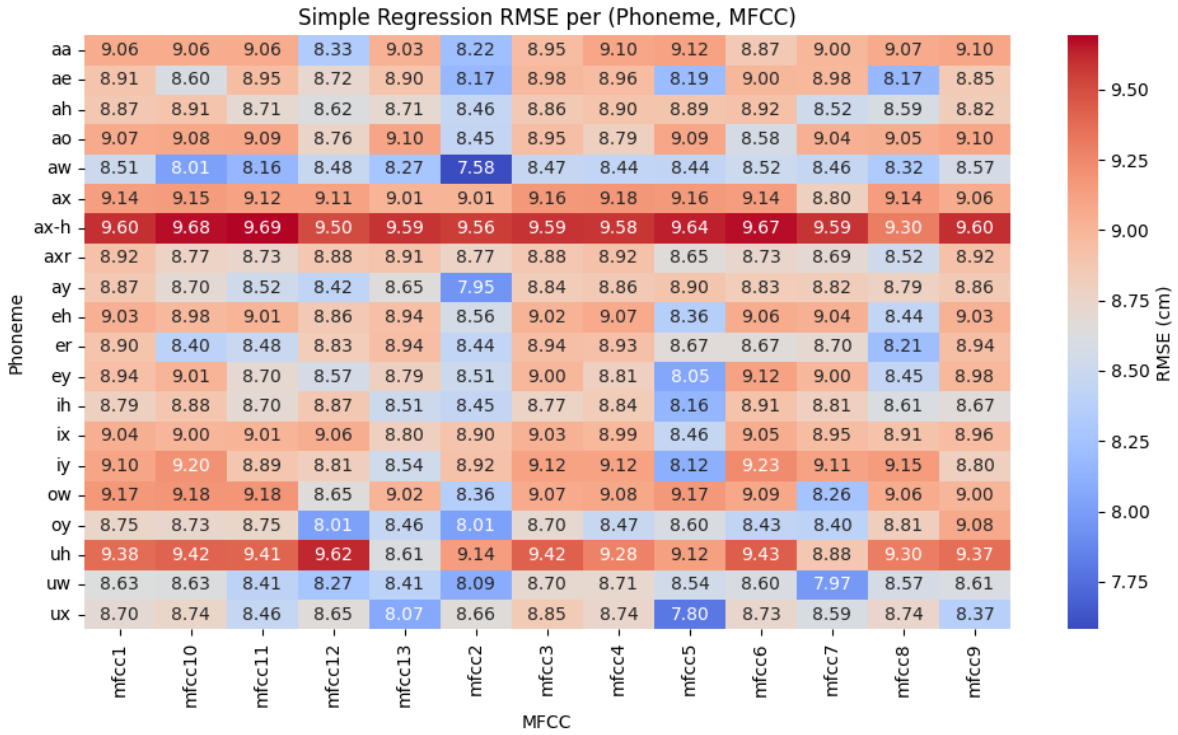


Figure 12: MFCCs RMSE per Phone per feature for Height Prediction Using Simple Linear Regression Model

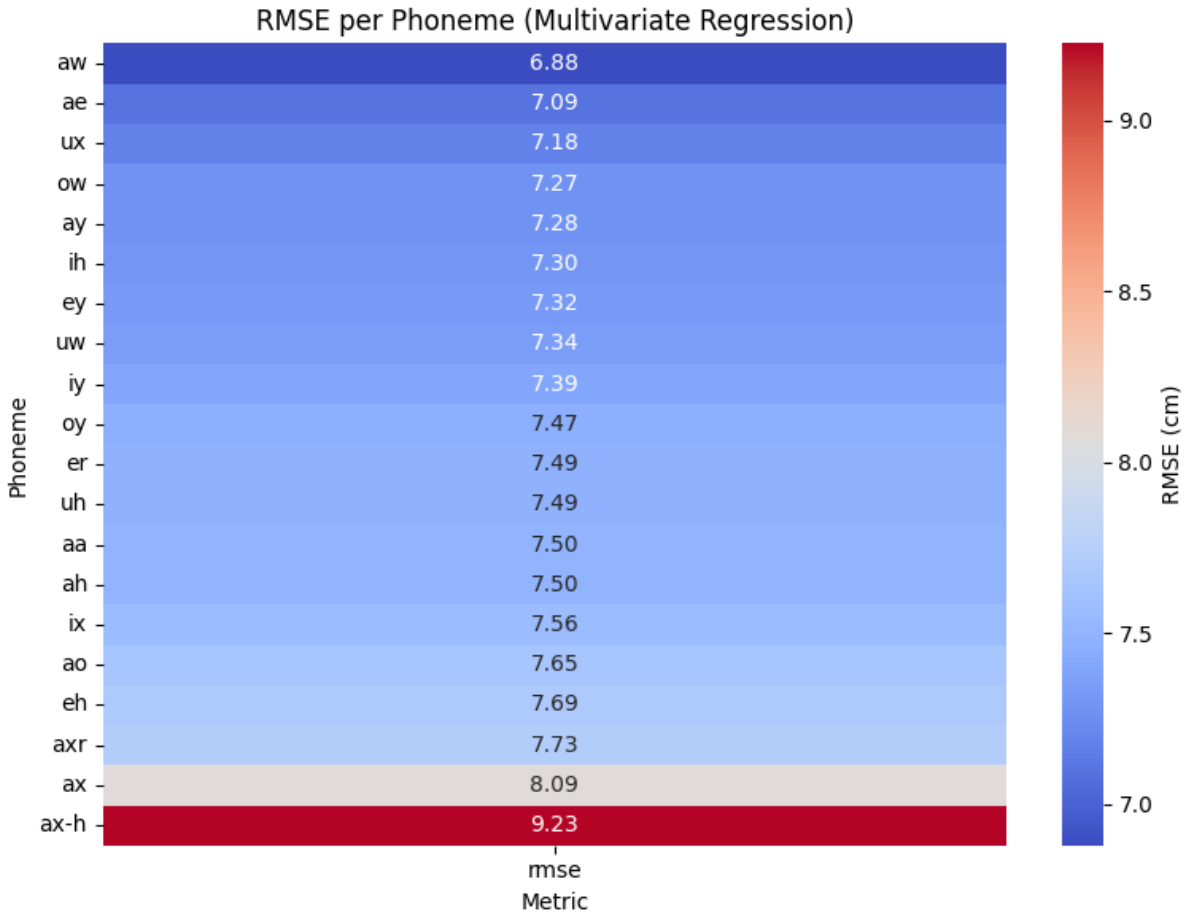


Figure 13: MFCCs RMSE per Phone for Height Prediction Using Multiple Linear Regression Model

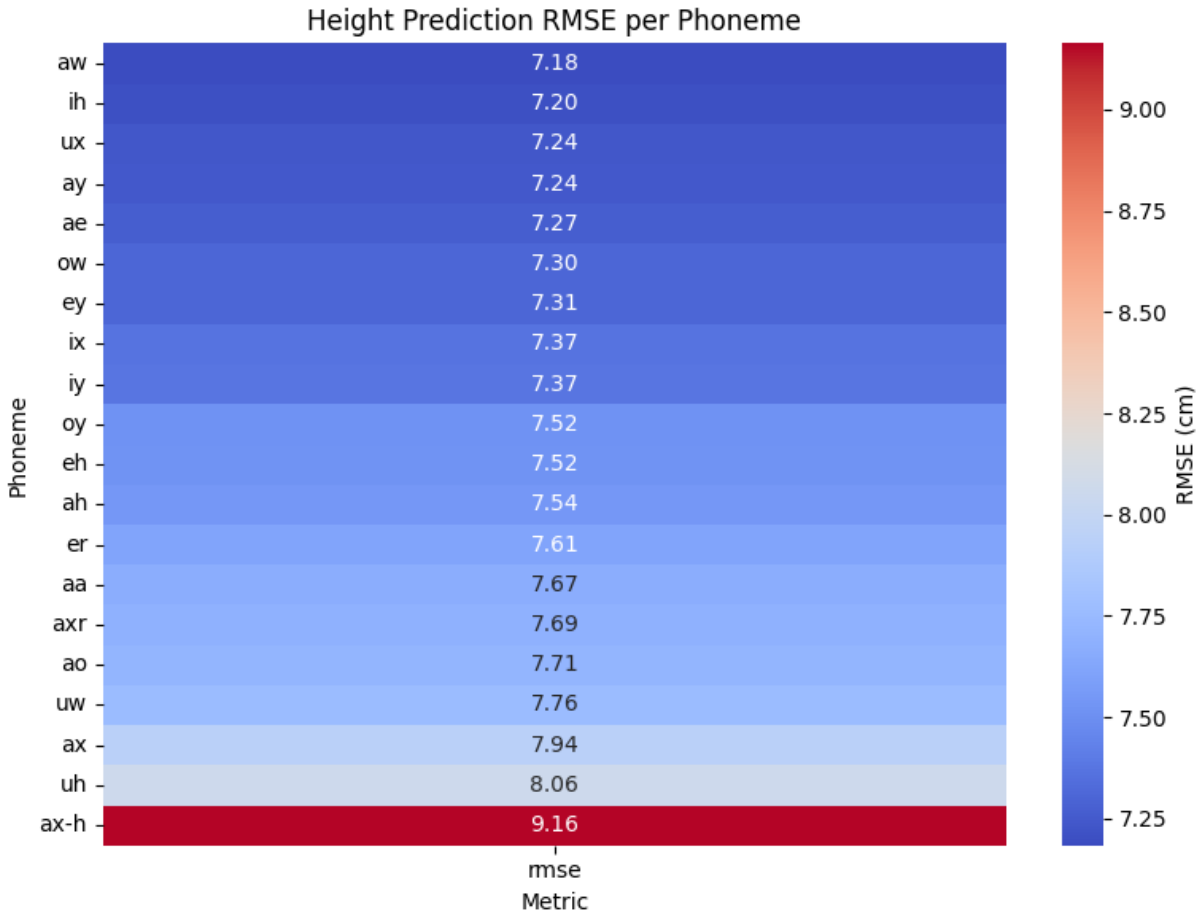


Figure 14: MFCCs RMSE per Phone for Height Prediction Using Random Forest Regression Model

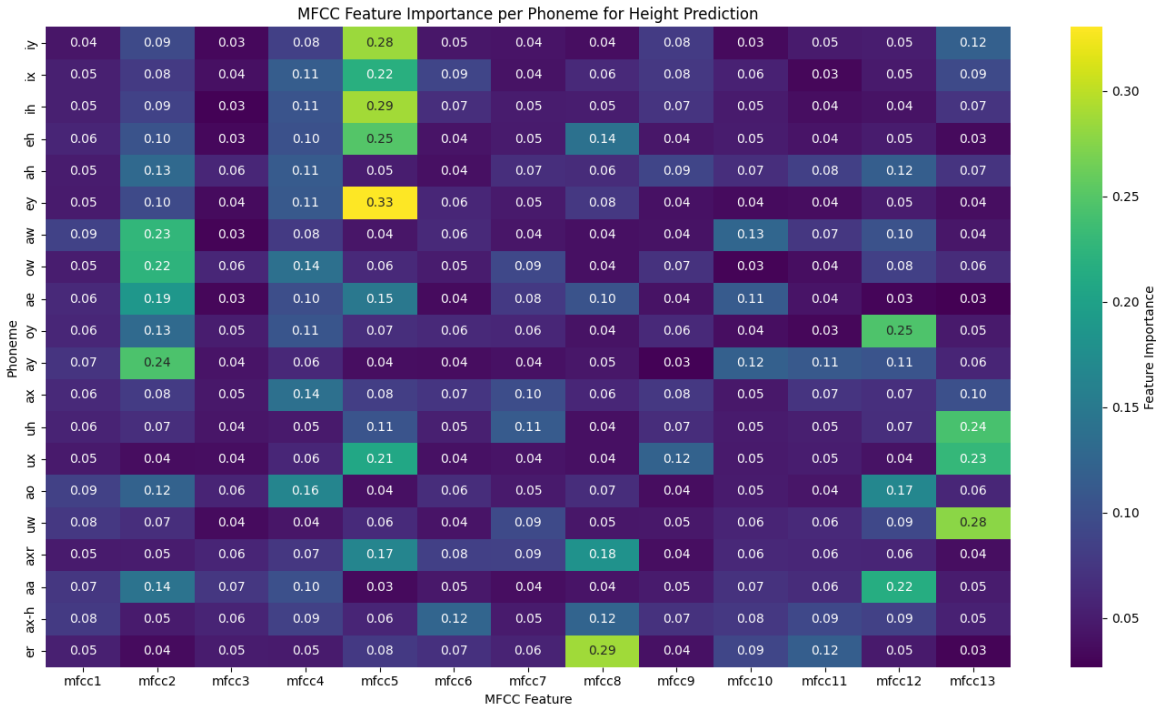


Figure 15: MFCCs Feature Importance per Phone per Feature for Height Prediction Using Random Forest Regression Model

B Visual Summary of Statistical Results

B.1 H1

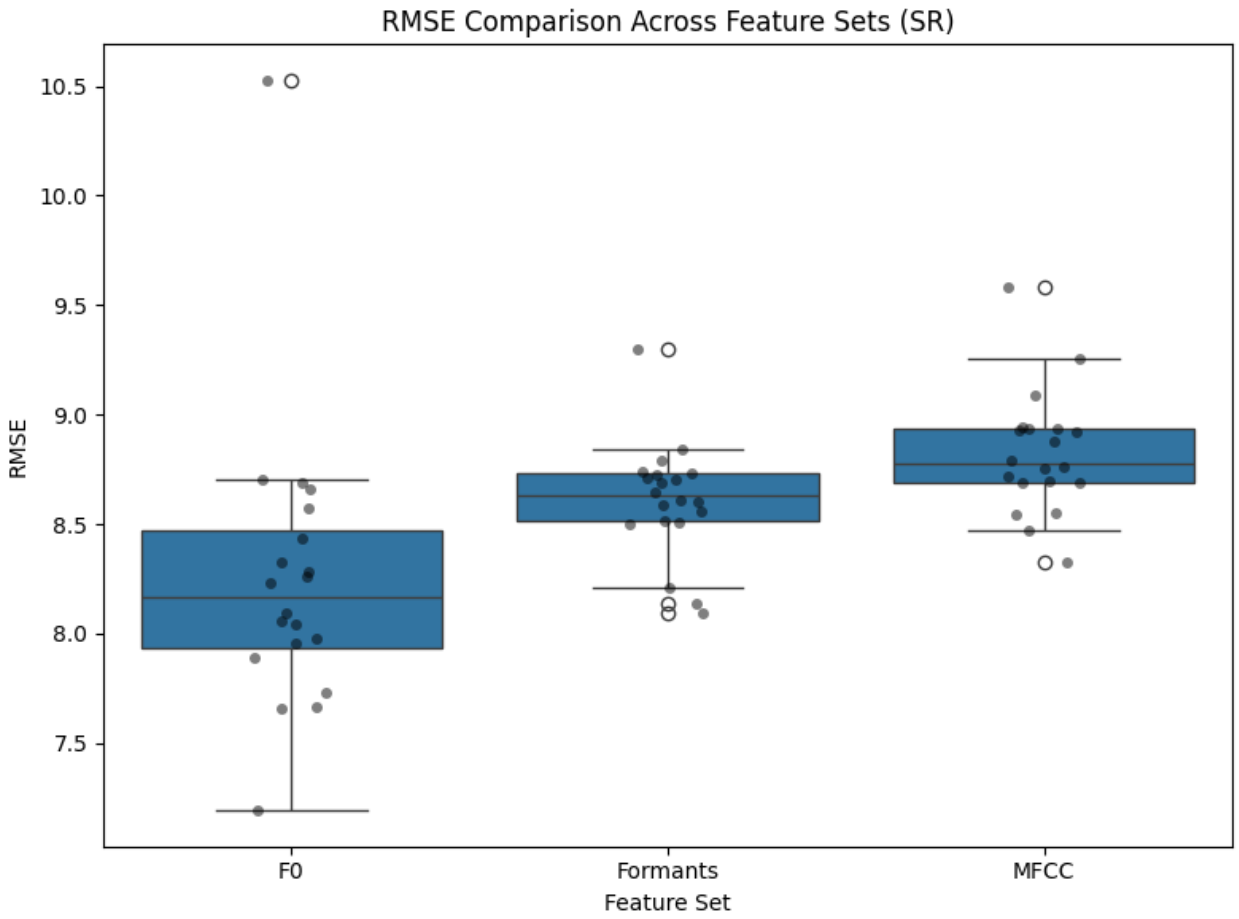


Figure 16: RMSE Comparison Across Formants and MFCC (Simple Linear Regression)

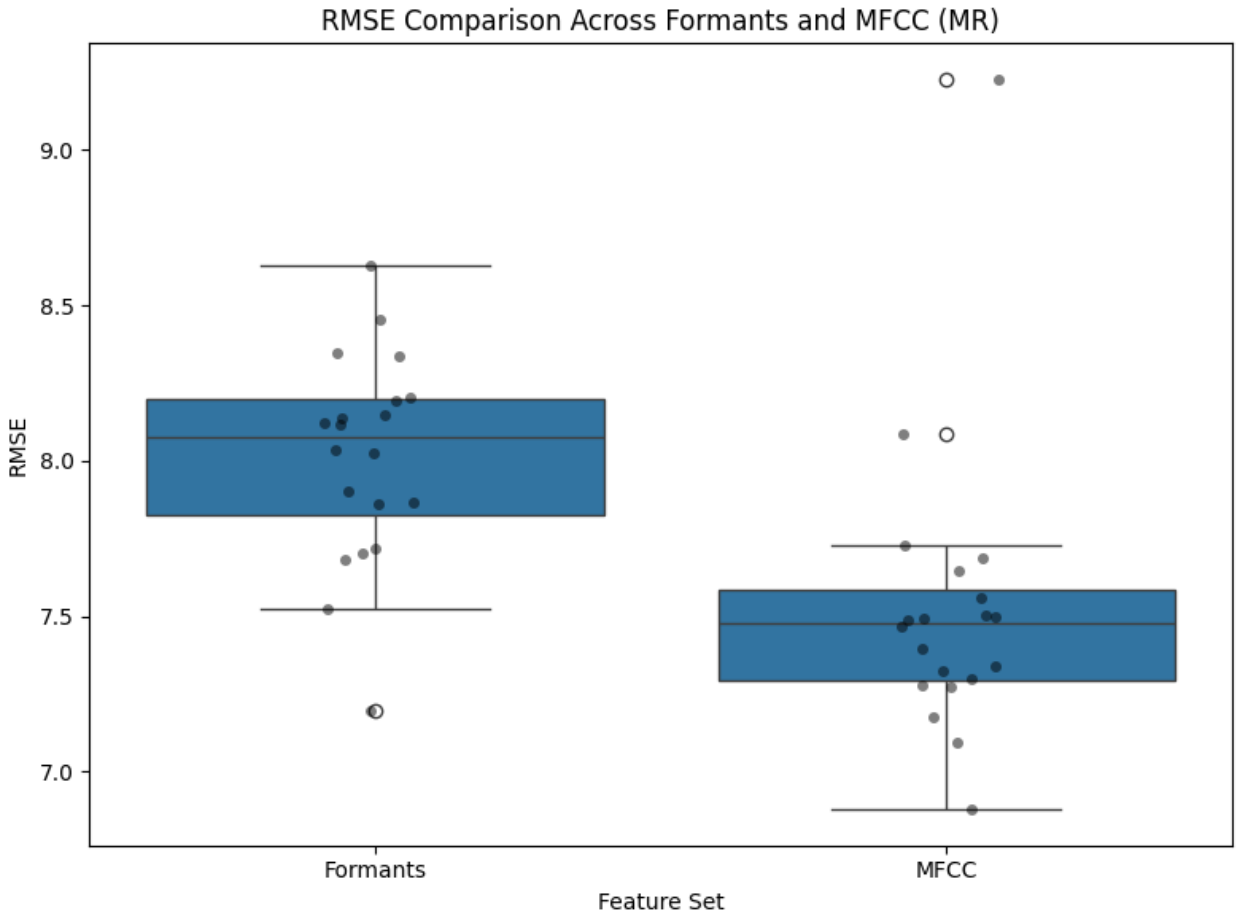


Figure 17: RMSE Comparison Across Formants and MFCC (Multiple Linear Regression)

B.2 H2

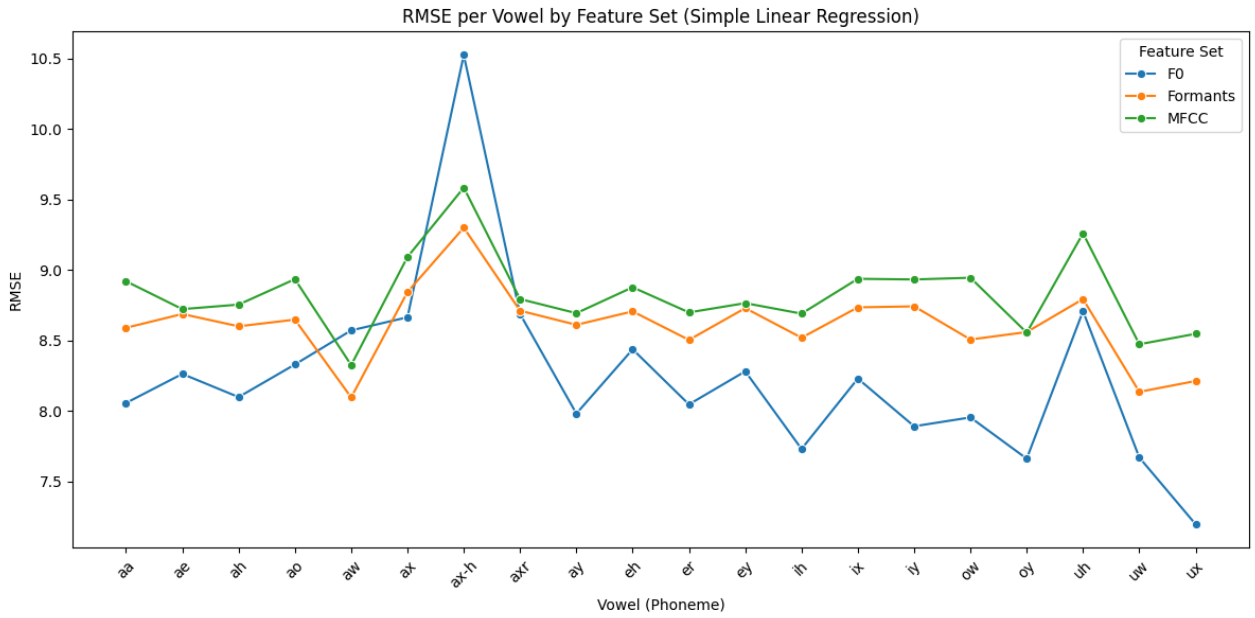


Figure 18: RMSE per Vowel by Feature Set (Simple Linear Regression)

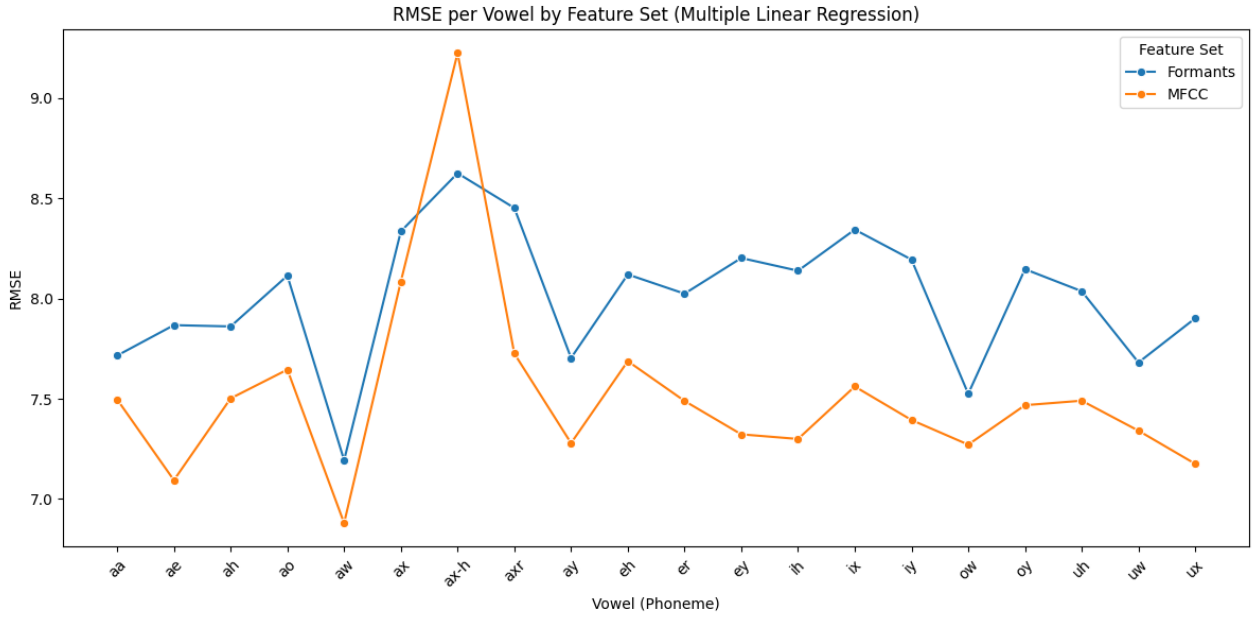


Figure 19: RMSE per Vowel by Feature Set (Multiple Linear Regression)

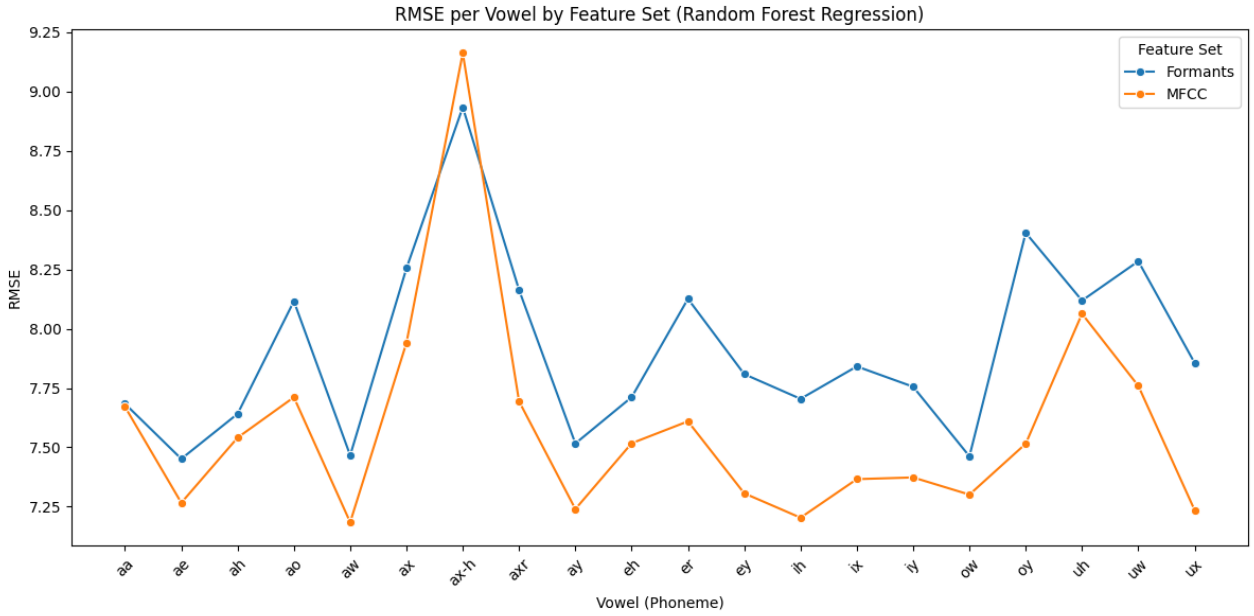


Figure 20: RMSE per Vowel by Feature Set (Random Forest Regression)

B.3 H3

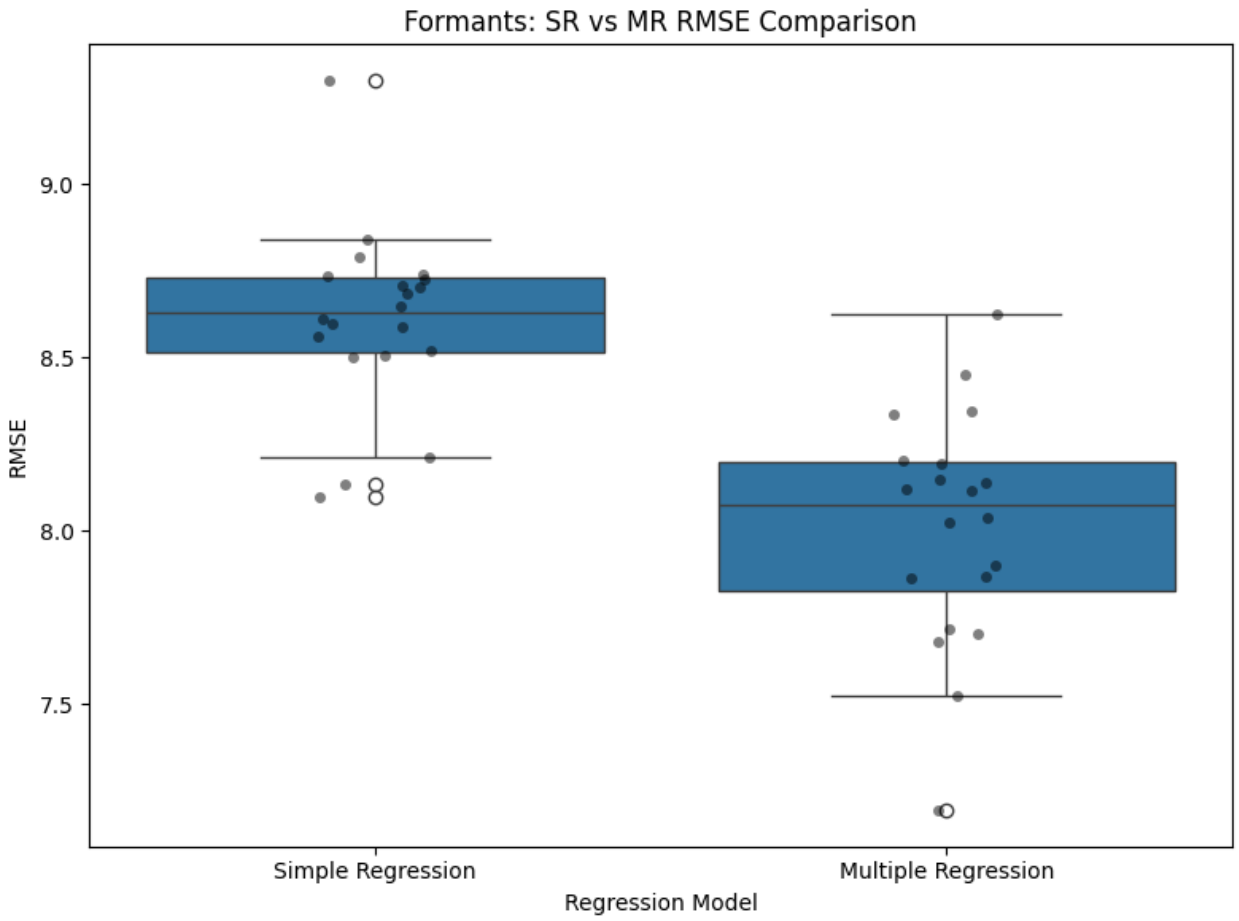


Figure 21: Formants: SR vs MR RMSE Comparison (SR vs MR)

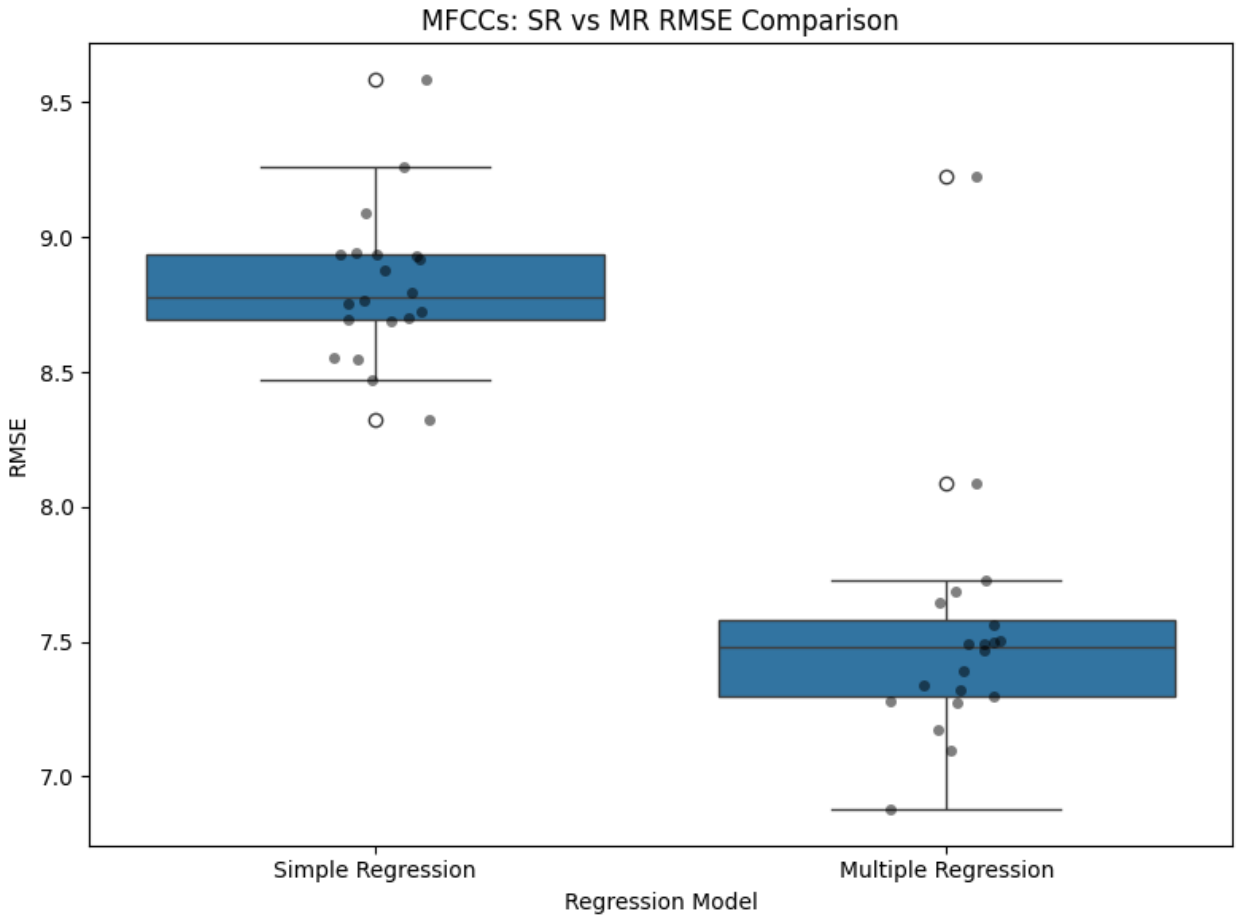


Figure 22: MFCCs: SR vs MR RMSE Comparison (SR vs MR)

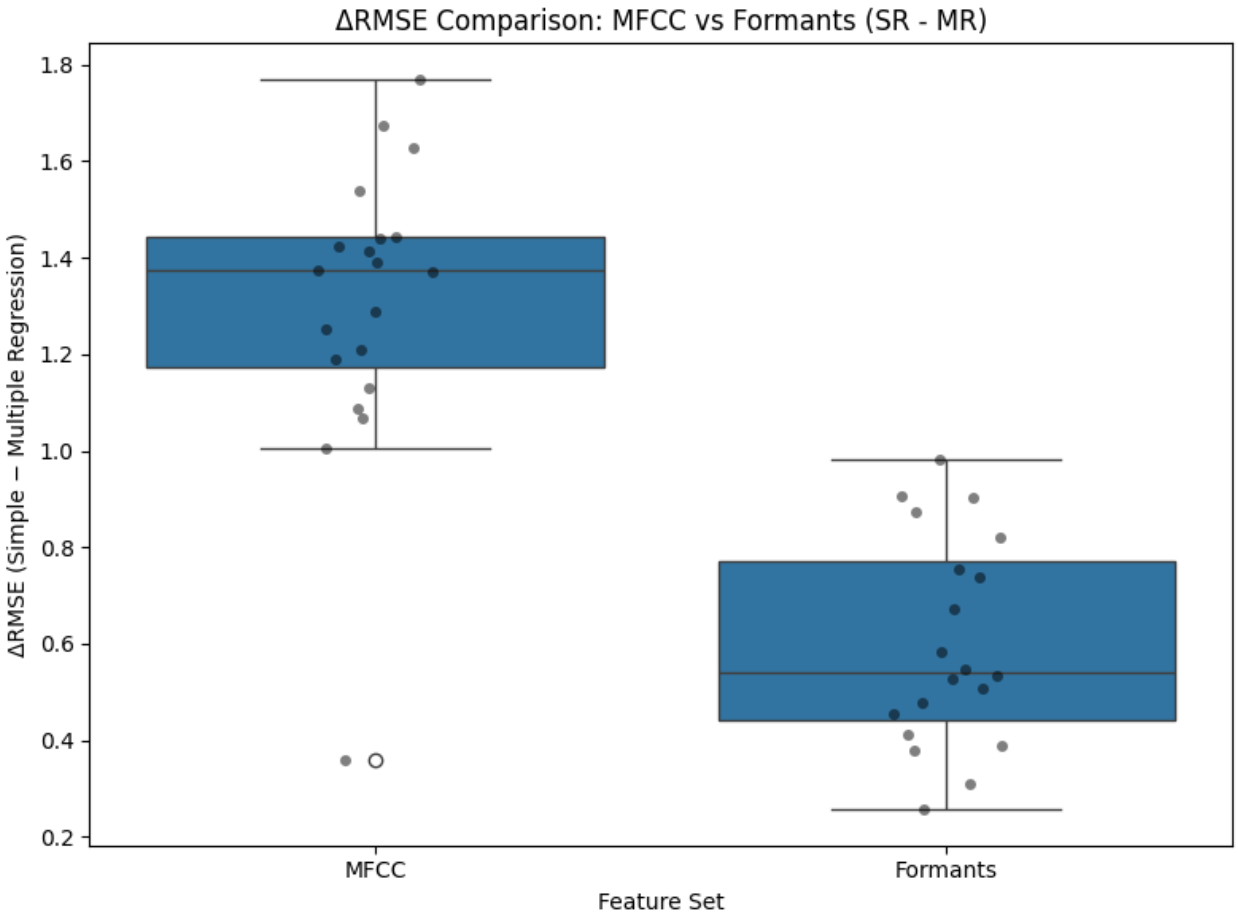


Figure 23: ΔRMSE Comparison: MFCC vs Formants (SR vs MR)

C Declaration of AI Use

Declaration I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used, this has been carefully acknowledged and referenced.

During the preparation of this thesis, I used ChatGPT-4o model for the following purposes:

1. Refining sentence structure, correcting grammar, and providing alternative lexicons across all chapters.
2. Assisting with LaTeX-compatible multi-row tables in Section 1 and 5, and equation formats in Section 3.
3. Providing guidance on the selection of statistical tests, including comparisons of alternative methods.
4. Supporting troubleshooting and debugging of Python code used for feature extraction, regression modelling, and statistical testing.

All content was subsequently reviewed, verified, and substantially modified by me.
Stella Siu 9 July 2025